

Review

Monocular Camera Localization in Known Environments: An In-Depth Review

Hailun Yan *, Albert Lau and Hongchao Fan

Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Høgskoleringen 1, 7034 Trondheim, Norway; albert.lau@ntnu.no (A.L.); hongchao.fan@ntnu.no (H.F.)

* Correspondence: hailun.yan@ntnu.no

Abstract

Monocular camera localization in known environments is a critical task for applications like autonomous navigation, augmented reality, and robotic positioning, requiring precise spatial awareness. Unlike localization in unknown environments, which builds maps in real time, this leverages pre-existing data for higher accuracy. This review comprehensively analyzes monocular camera localization methods in known environments, categorizing them into 2D-2D feature matching, 2D-3D feature matching, and regression-based approaches. It consolidates foundational techniques and recent advancements, providing inter-class and intra-class performance comparisons on mainstream datasets. Key findings show that 2D-3D methods generally offer the highest accuracy, especially in structured outdoor environments, due to robust use of 3D spatial information. However, recent scene coordinate regression methods, such as ACE and ACE++, achieve comparable or superior performance in indoor scenes with more efficient pipelines. This review highlights challenges and proposes future directions: (1) synthetic data generation to meet deep learning demands, while addressing domain gaps; (2) improving generalization to unseen scenes and reducing retraining; (3) multi-sensor fusion for enhanced robustness; (4) exploring transformer-based and graph neural network architectures; (5) developing lightweight models for real-time performance on resource-constrained devices. This review aims to guide researchers and practitioners in method selection and identify key research directions.

Keywords: monocular camera localization; camera pose estimation; feature matching-based localization; regression-based localization; deep learning

Academic Editor: Carlos Lima

Received: 25 January 2026

Revised: 16 February 2026

Accepted: 24 February 2026

Published: 27 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Overview

In recent years, monocular camera localization has emerged as a cornerstone technology for a wide array of applications, ranging from autonomous navigation and robotics to augmented reality and urban planning. Monocular camera localization involves estimating the precise 6 Degrees of Freedom (6-DoF) pose of a monocular camera using 2D visual information, which is crucial for tasks that require spatial awareness and interaction with the physical world. Compared to localization systems that rely on other sensor data, such as Light Detection and Ranging (LiDAR) scans, Radio Detection and Ranging (RADAR) scans, RGB-D images, and wireless signals such as WiFi, monocular-camera-

based localization offers the most flexibility and cost-effectiveness. Companies like Tesla, for example, aim to achieve full autonomy in their self-driving cars using only 2D cameras, motivated by the lower cost of cameras compared to LiDAR and recent advancements in deep learning technologies that can leverage the rich visual information provided by these images like never before [1–4]. Furthermore, the ubiquity of cameras, embedded in everyday devices such as mobile phones, expands the potential applications of monocular camera localization, making it accessible and practical for widespread use across diverse contexts.

While the problem of monocular camera localization has been extensively studied, it is often divided into two distinct scenarios: localization in known environments, where prior knowledge of the environment exists, and localization in unknown environments, where such prior knowledge is absent. Localization in known environments presents unique advantages over its counterpart in unknown settings. In environments where the structure, features, and layout are pre-determined and well documented, localization can achieve higher accuracy, greater computational efficiency, and faster convergence times. This is particularly important in applications like train positioning and automated warehouse management, where the target operates within a fixed and predictable space, or for self-driving cars that find themselves in tunnels or underground parking lots where GNSS is compromised. The use of pre-existing data allows for more robust and reliable localization, reducing the need for intensive real-time computation and mapping, as required by techniques based on Simultaneous Localization and Mapping (SLAM), which are tailored for unknown environments. This review focuses on monocular camera localization in known environments.

1.2. Basic Principles of Monocular Camera Localization

Monocular camera localization is the task of estimating the 6-DoF pose of a single camera from one RGB image. The 6-DoF pose consists of the camera's 3D position (translation: x, y, z) and its 3D orientation (usually represented by a 3×3 rotation matrix, R , or a unit quaternion).

The camera is modeled using the standard pinhole projection model. A 3D point, $\mathbf{X}_w = [X_w, Y_w, Z_w]^T$, in the world coordinate system is projected onto the 2D image plane as pixel coordinates (u, v) according to the following relationship:

$$s \times [u, v, 1]^T = K \times [R|t] \times \mathbf{X}_w \quad (1)$$

where K is the 3×3 intrinsic calibration matrix:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},$$

f_x and f_y are focal lengths in pixels, c_x and c_y are the principal point, s is a positive scale factor, R is the 3×3 rotation matrix, and t is the 3×1 translation vector.

Two fundamentally different scenarios exist:

1. Localization in unknown environments (online SLAM): The system builds a map of the scene in real time while simultaneously estimating the camera's pose.
2. Localization in known environments (the focus of this survey): A pre-existing map or reference dataset of the environment is already available. The map can take several forms:
 - i. A database of reference images with known camera poses;
 - ii. A 3D point cloud or mesh with associated visual descriptors;
 - iii. A learned neural scene representation.

In the known-environment case, the task generally involves relating the query image to the prior map to recover the unknown camera pose $(\mathbf{R}, t)$. This can be achieved through explicit feature correspondences or implicit learned mappings.

Methods in this survey typically follow a high-level workflow, though variations exist (e.g., regression-based methods may integrate or bypass certain steps for end-to-end efficiency):

- (1) Feature extraction from the query image (using handcrafted keypoints and descriptors, learned local features, or dense features from CNNs/Transformers; in regression methods, this is often implicit within the network).
- (2) Relating the query to the known environment (via explicit matching, like 2D-2D image-to-image or 2D-3D image-to-point-cloud correspondences, or implicitly through neural regression of poses or scene coordinates).
- (3) Pose computation either by solving a geometric problem (e.g., essential-matrix decomposition for 2D-2D; Perspective-n-Point solver for 2D-3D) or by direct neural regression (e.g., outputting the 6-DoF pose or per-pixel 3D coordinates followed by a lightweight solver).

1.3. Existing Surveys and Limitations

Several review papers have contributed to the field of monocular camera localization. For example, Wu et al. [5] present a comprehensive overview of image-based camera localization, categorizing approaches based on whether the environment is known or unknown, and discuss Perspective-n-Point (PnP) solutions for known environments and Simultaneous Localization and Mapping (SLAM) for unknown environments. Piasco et al. [6] focus on the benefits of heterogeneous data in visual-based localization (VBL), emphasizing how various types of input, such as point clouds and semantic information, enhance robustness against environmental changes. Xin et al. [7] provide a detailed review of VBL methods by classifying them into image-based, learning-based, and 3D structure-based localization, highlighting their strengths and limitations in complex scenes. Humenberger et al. [8] explored the role of image retrieval in visual localization through an exhaustive benchmark, categorizing localization paradigms into pose approximation, local 3D map-based localization, and global 3D map-based localization. They emphasize the need for retrieval approaches tailored to specific localization paradigms and note limitations in current retrieval methods for dynamic and blurred scenes.

Despite these valuable contributions, existing review studies are either broad in scope or lack detailed analyses and comparisons of methods and datasets. In addition, many studies focus heavily on methods applicable to both known and unknown environments without delving into the nuances of pre-mapped environments. Furthermore, given the rapid advancements in computer vision, deep learning, and sensor technology, there is a pressing need for an updated review that reflects the current state of the field. This paper aims to fill this gap by providing an in-depth review focused specifically on monocular camera localization in known environments. It consolidates both foundational techniques and recent developments to guide practitioners and researchers in understanding and selecting appropriate methods and identifying emerging trends and challenges within this specialized domain.

1.4. Classification of Methods

As shown in Figure 1, we broadly categorize the methods for monocular camera localization in known environments into three main types based on their underlying principles: 2D-2D feature matching-based localization, 2D-3D feature matching-based localization, and regression-based localization. The 2D-2D matching-based methods determine the camera pose by comparing a query image against reference images within a pre-

constructed pose-tagged database. This category is further divided into approximate 2D-2D localization, which estimates a coarse camera pose based on nearest-neighbor image retrieval, and precise 2D-2D localization, which refines the initial estimation using relative pose estimation techniques to achieve high accuracy. In contrast, 2D-3D feature matching-based localization relies on matching 2D features from the query image to 3D scene points in a pre-built 3D map to compute the exact 6-DoF camera pose. This approach is divided into direct 2D-3D localization, where direct matching between image features and 3D points occurs, and hierarchical 2D-3D localization, which reduces computational overhead by narrowing down the search space through an initial rough pose estimation. Lastly, regression-based localization bypasses explicit feature matching by leveraging deep learning models to directly predict the camera's pose. This category includes absolute pose regression, which estimates the global 6-DoF pose, and scene coordinate regression, which predicts the 3D scene coordinates for each pixel in the query image and refines the pose using a PnP solver.

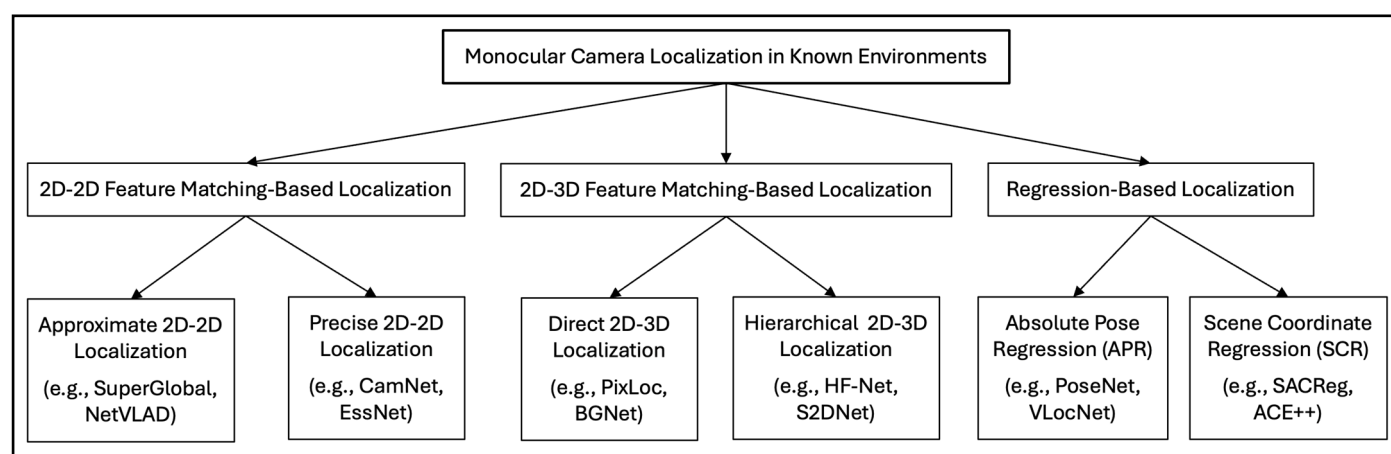


Figure 1. Classification of monocular camera localization methods for known environments.

1.5. Survey Structure

The remainder of this paper is organized as follows. Section 2 provides an overview of 2D-2D feature matching-based localization methods, discussing both approximate and precise approaches. Section 3 presents 2D-3D feature matching-based localization methods, including the creation of 3D prior maps, and direct and hierarchical strategies. Section 4 reviews regression-based localization techniques, covering absolute pose regression and scene coordinate regression methods. Section 5 reviews the datasets and evaluation metrics commonly used to benchmark localization methods, summarizing their metadata and unique characteristics. Section 6 conducts a comparative analysis of the reviewed methods by evaluating their performance across key benchmark datasets. The analysis considers inter-class comparison, which evaluates how different localization paradigms—2D-2D matching, 2D-3D matching, and regression-based methods—perform relative to one another, and intra-class comparison, which examines variations within each category. Section 7 discusses key challenges and proposes future research directions for monocular camera localization in known environments. Finally, Section 8 concludes the paper by summarizing key insights and offering actionable recommendations for researchers and practitioners.

2. 2D-2D Feature Matching-Based Localization

2D-2D feature matching-based localization (2D-2D localization) is a widely studied field in monocular camera localization in known environments. This approach primarily

involves matching a query image against a pre-constructed database of pose-tagged reference images by comparing their visual features. After retrieving the most similar image or set of images from the database, the system determines the posture information about the query image relative to the retrieved database images. An illustration of 2D-2D feature matching is shown in Figure 2. Based on a trade-off between localization accuracy and computational efficiency, 2D-2D localization techniques can be categorized into two sub-types: Approximate 2D-2D localization and Precise 2D-2D localization. A flowchart illustrating the general workflow of these methods is shown in Figure 3.

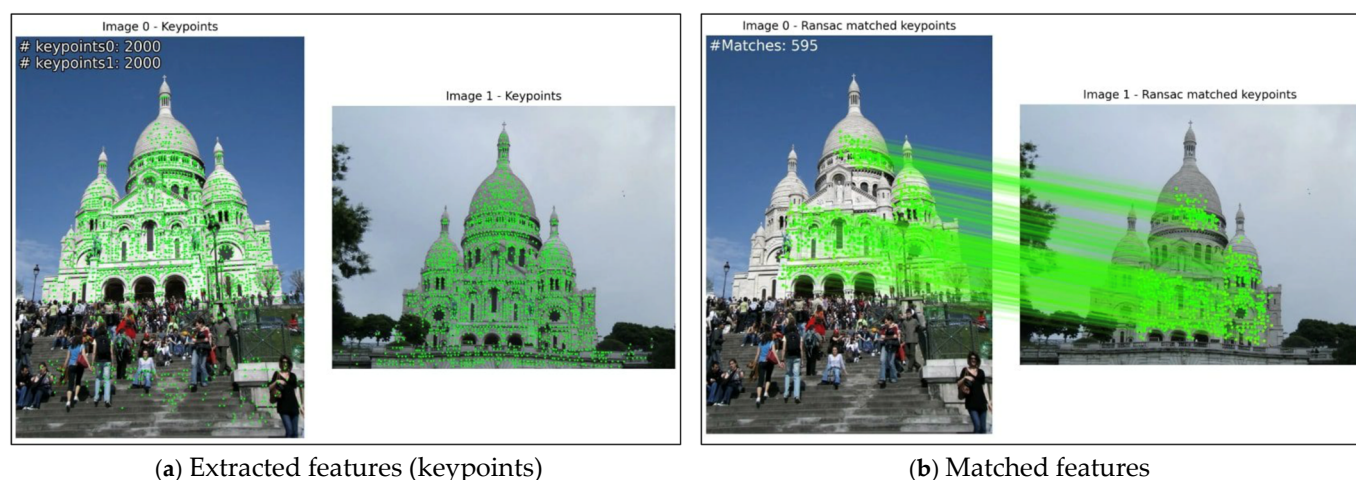


Figure 2. An example of feature matching in 2D-2D methods. (a) Features are extracted from a query image and a database image, shown in green dots. (b) Correspondences between the two images are established by matching high-similarity feature pairs, which are connected by green lines.

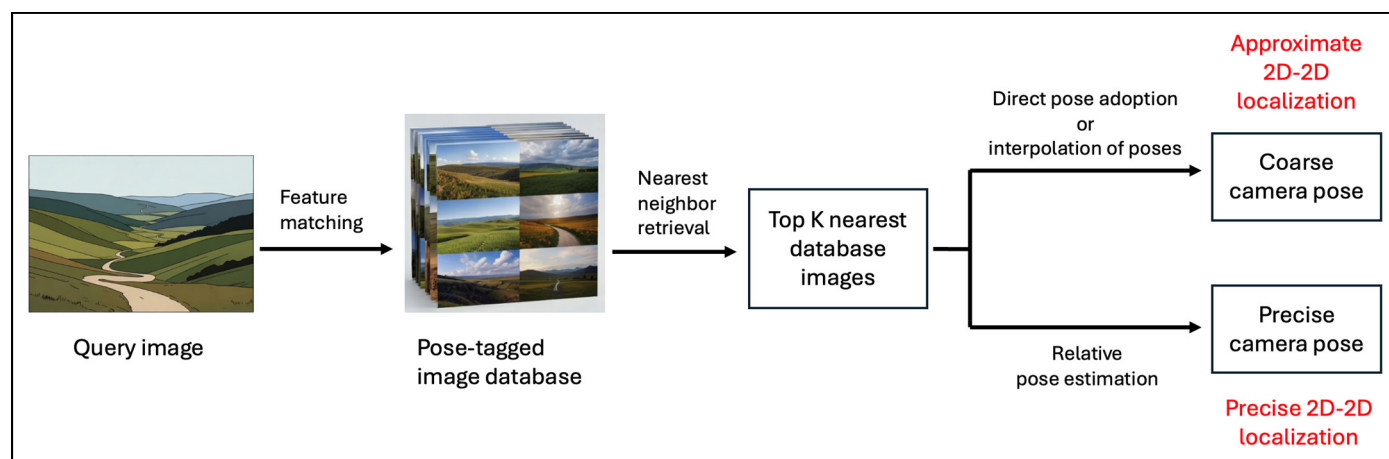


Figure 3. Overview of 2D-2D feature matching-based localization methods. (The top branch belongs to approximate 2D-2D localization, while the bottom branch belongs to precise 2D-2D localization.)

2.1. Approximate 2D-2D Localization

Approximate 2D-2D localization, also known as place recognition or visual geo-localization, typically approximates a rough 6-DoF pose for the query image by adopting the pose of the nearest neighbor or interpolating (linearly) between the poses of the top k retrieved images. This approach enables efficient localization at the cost of accuracy in scenarios where coarse camera poses are sufficient. The following is an overview of 2D-2D feature matching-based localization methods.

Conventional Approximate 2D-2D localization methods primarily rely on hand-crafted feature descriptors to extract local key point features from query and reference images. These key points are used to measure image similarity, enabling the system to retrieve the closest match within a reference database. However, modern techniques have since incorporated deep learning-based techniques, which capture more complex and richer image-level features, leading to superior performance over traditional methods. To provide a thorough understanding of the evolution of Approximate 2D-2D localization techniques, we present a chronological review of the representative methods, covering both the early conventional and recent deep learning-based approaches.

2.1.1. Conventional Methods

Foundational visual word-based matching techniques. The early successes of Approximate 2D-2D localization in retrieving images from large-scale databases were primarily achieved by utilizing rather mature text retrieval architectures, particularly the bag-of-words (BoW) model. This adaptation was driven by the need to handle large-scale image databases efficiently, much like text retrieval systems manage vast collections of documents.

In text retrieval, documents are represented as unordered collections of words, disregarding grammar and word order—hence the term “bag-of-words.” Each document is then represented by a histogram indicating the frequency of each word from a predefined vocabulary. This representation allows for efficient indexing and retrieval using techniques like inverted files. Similarly, in image retrieval, an image can be considered analogous to a document, and local image features can be considered analogous to words. By treating local features as visual words, images can be represented as histograms over a visual vocabulary, enabling the application of text retrieval methods to images.

The foundational concept of representing objects using visual words originates with the “Video Google” approach introduced by Sivic [9]. This method was inspired by text retrieval systems and laid the groundwork for using viewpoint-invariant region descriptors with visual word indexing. It efficiently matches images through vector quantization, but it suffers from issues such as missing relevant matches due to quantization errors. Despite this limitation, the core idea of visual words became a cornerstone for subsequent image retrieval methods. Building upon this foundation, Nister and Stewenius [10] introduced the vocabulary tree method, which addressed the scalability of visual word-based systems. By organizing local descriptors into a hierarchical tree, their system improved retrieval efficiency and memory usage. This innovation allowed for logarithmic-time retrieval and real-time adaptability, making it suitable for real-time applications like vision-based SLAM. The vocabulary tree was a key development in enabling large-scale object recognition and image retrieval.

Enhanced spatial matching and query refinement. Advancing the precision of image retrieval, Philbin et al. [11] introduced spatial verification (SP) to complement visual word matching. By ensuring the spatial arrangement of features between query and retrieved images was consistent, they mitigated false positives caused by visual word ambiguity. This technique significantly enhanced the robustness of image retrieval. That same year, Chum et al. [12] refined the retrieval process further through query expansion (QE). By iteratively refining the query using spatially verified matches, this method enhanced recall and precision, especially in datasets with occlusions. Both techniques played a key role in early visual word-based approaches by eliminating false positives and increasing retrieval robustness.

Probabilistic models. A departure from strict visual word matching came with the introduction of FAB-MAP [13], which leveraged probabilistic modeling. FAB-MAP improved place recognition by modeling co-occurrence relationships between visual words,

making it more effective in handling perceptual aliasing—where distinct locations look similar. This probabilistic approach introduced an advanced level of robustness to place recognition, particularly in dynamic environments.

Efficient feature representations. Moving toward more compact and efficient representations, Perd'och et al. [14] proposed a compact, discretized representation of elliptical regions, learned through minimizing the average reprojection error in the space of ellipses, which reduces memory usage to 24 bits per feature while maintaining high retrieval performance. This advancement tackled the growing memory demands of large-scale retrieval systems. Jégou et al. [15] introduced VLAD (Vector of Locally Aggregated Descriptors). By aggregating local image descriptors into a compact vector, VLAD outperformed traditional bag-of-words approaches while using less memory. This method allowed for efficient large-scale retrieval, requiring as few as 16 to 32 bytes per image, marking a significant leap in the practicality of image retrieval for large datasets.

Innovating query expansion. Chum et al. [16] proposed a novel approach to query expansion by introducing incremental Spatial Re-Ranking (iSP) and contextual Query Expansion (ctx QE). The method first applies spatial verification to eliminate false positives in the initial results, ensuring that only reliable images contribute to query refinement. Using iSP, it incrementally updates the query model by incorporating verified images, while ctx QE expands the query by learning spatially consistent features from nearby image regions. This hybrid approach improves both recall and precision by iteratively refining the query.

Refinements in descriptor matching and database augmentation. Arandjelović and Zisserman [17] introduced three significant enhancements to image retrieval systems. First, they proposed RootSIFT, a simple transformation of the SIFT descriptor using the Hellinger kernel, which substantially improved retrieval performance without additional computational cost. Second, they introduced Discriminative Query Expansion (DQE), which used a linear SVM to refine query expansion, outperforming traditional methods by learning to re-rank images based on relevance. Lastly, they introduced spatial database-side feature augmentation (SPAUG) by incorporating spatial verification, ensuring that only spatially consistent features were used for augmentation. These advancements raised the bar for performance on multiple image retrieval benchmarks.

Handling repetitive structures. The challenge of repetitive structures, such as building façades, fences, and road markings, was addressed by Torii et al. [18]. Their technique modified visual word weighting to prevent over-counting in the presence of repetitive patterns, improving matching accuracy in structured environments. However, manual adjustments were necessary for this approach to function optimally.

Feature aggregation for discriminative representations. Kim et al. [19] introduced Per-Bundle VLAD (PBVLAD), which aggregates multiple SIFT features detected within maximally stable extremal regions into a fixed-size vector for more effective and discriminative feature representation. By focusing only on the most relevant features, PBVLAD reduced false positives and enhanced retrieval accuracy, outperforming traditional VLAD and SIFT-based methods. This refinement was crucial for improving localization performance in large-scale environments.

View synthesis and DenseVLAD. To further enhance robustness of image retrieval in significant changes in appearance, such as illumination, seasons, and structural modifications, Torii et al. [20] introduced a view synthesis method and DenseVLAD descriptors. The authors synthesized virtual views from Google Street View panoramas and approximated depth maps, which are aligned with query images captured from similar viewpoints. In addition, they proposed DenseVLAD descriptors, which leverage densely sampled local gradient-based descriptors over a fixed grid instead of traditional keypoints

for improved robustness. DenseVLAD offers robustness against keypoint detection failures and captures both fine-grained details and global context.

SMK and ASMK. Tolias et al. [21] introduced the Selective Match Kernel (SMK) technique to improve local feature matching. SMK prioritizes strong matches between local descriptors while suppressing weaker, less reliable matches. This technique addresses the burstiness phenomenon, where certain features dominate the matching process, leading to more balanced and accurate retrieval results. Built upon the SMK, the authors also proposed Aggregated Selective Match Kernel (ASMK), which aggregates local descriptors assigned to the same visual word. This joint encoding of local descriptors not only enhances robustness against descriptor redundancy but also improves scalability by reducing memory requirements. ASMK further handles burstiness through advanced normalization techniques, making it more effective for large-scale image retrieval tasks.

In summary, the conventional Approximate 2D-2D localization methods—from early bag-of-words and vocabulary tree approaches to advanced spatial verification, query expansion, and VLAD/ASMK aggregations—established the core principles of efficient, scalable image retrieval using handcrafted local descriptors. Although effective, these methods rely on handcrafted features and fixed models, which inherently limit their adaptability to varying scene conditions and complexities, particularly when compared with the deep learning methods that emerged later.

2.1.2. Deep Learning-Based Methods

In later years, the rise of deep learning revolutionized the field of computer vision, providing a more powerful and flexible framework for approximate 2D-2D localization. Deep learning-based methods utilize deep convolutional neural networks (CNNs) to automatically learn and extract robust feature representations from images. These deep features have been shown to outperform handcrafted descriptors by capturing higher-level semantic information, making them better suited for complex and diverse image retrieval tasks.

Early CNN-based descriptors. The initial breakthroughs in deep learning for image retrieval came with the introduction of global descriptors like SPoC (Sum-Pooling of Convolutional Features) by Babenko and Lempitsky [22], which aggregates features from the last convolutional layer of a pre-trained CNN using simple sum pooling. This simple yet effective approach demonstrated that deep convolutional features could outperform traditional handcrafted features like SIFT, paving the way for CNN-based methods in image retrieval. Around the same time, Tolias et al. [21] introduced R-MAC (Regional Maximum Activations of Convolutions), a technique that extracts regional features at various scales, encoding multiple image regions for more precise retrieval. This innovation demonstrated the potential of CNNs to handle image variations more effectively than global descriptors alone. Building on these early methods, Gordo et al. [23] proposed Deep Image Retrieval (DIR), which enhances R-MAC by introducing a triplet ranking loss to improve the quality of regional features and replacing the rigid grid with a more flexible Region Proposal Network (RPN). This allows for more content-aware pooling and better localization of regions of interest. Also in 2016, Arandjelovic et al. [24] introduced NetVLAD, which integrated VLAD pooling into a CNN architecture, enabling end-to-end learning for place recognition under extreme variations in viewpoint, lighting, and seasonal conditions. Meanwhile, Radenović et al. [25] introduced the MAC (Maximum Activations of Convolutions) model, which provides a compact image representation derived from the maximum activation values of feature maps in a convolutional neural network. By applying max-pooling over each feature map, MAC produces a fixed-length descriptor that is highly efficient for image retrieval tasks. It eliminates the need for manual region selection, leveraging the strongest activations to represent the most salient image features. The MAC descriptor is

further enhanced with L₂-normalization for robust similarity measurement. Furthermore, Kalantidis et al. [26] introduced the CroW (Cross-dimensional Weighting) model to aggregate deep convolutional features for image retrieval. It applies non-parametric spatial and channel-wise weighting to enhance the relevance of highly active spatial regions and reduce the impact of bursty features in convolutional layers. CroW improves upon traditional feature aggregation methods by addressing feature imbalance and preserving discriminative information.

Attention mechanism. In 2017, Noh et al. [27] introduced DELF (DEep Local Features), which incorporates attention mechanisms for local feature selection. DELF uses an attention mechanism to select semantically relevant keypoints, improving the discriminative power of local features. This approach allows DELF to excel in challenging environments with clutter, occlusion, and variable scales.

E2E-R-MAC. Gordo et al. [28] introduced an end-to-end learning framework for image retrieval, building upon their previous R-MAC descriptor. It integrates all stages of descriptor extraction, including region pooling and dimensionality reduction, into a single differentiable CNN. The model is trained using a triplet-based ranking loss to optimize the retrieval task directly, rather than relying on classification. Improvements include a learned pooling mechanism and multi-resolution descriptors, enabling robust retrieval across scales and viewpoints. E2E-R-MAC significantly outperformed state-of-the-art methods on benchmarks like Oxford5K and Paris6K at the time while remaining computationally efficient and scalable.

GeM and α QE. In 2018, Radenović et al. [29] introduced a novel approach that leveraged SfM to automatically generate training data, making it easier to fine-tune CNNs for image retrieval. Their work also introduced the Generalized Mean (GeM) pooling layer, a trainable pooling method that generalizes max and average pooling, significantly improving feature aggregation and retrieval performance. This was combined with learned whitening, a discriminative whitening technique, to optimize descriptor representations further. Additionally, the study proposed α -weighted Query Expansion (α QE) as an improvement over the standard Average Query Expansion (AQE). While AQE aggregates descriptors of top-ranked images with equal weights, α QE assigns weights to these images based on their similarity to the query, providing a more robust and stable query expansion strategy.

Improving the ASMK. Teichmann et al. [30] proposed a variant of the ASMK framework named Regional Aggregated Selective Match Kernel (R-ASMK) to enhance image retrieval. Unlike traditional ASMK, which aggregates descriptors from the entire image, R-ASMK leverages deep learning-based region detection to aggregate local features selectively from relevant parts of an image, improving discriminative power and down-weighting irrelevant features. This method retains the dimensionality of the image descriptor while improving accuracy.

Hypergraph model for enhanced spatial propagation. An et al. [31] introduced a novel Hypergraph Propagation + Community Selection model to address the limitations of traditional query expansion and diffusion methods. The proposed Hypergraph Propagation efficiently propagates spatial information across a hypergraph structure that connects local features across multiple images. This approach achieves better retrieval accuracy by resolving ambiguities caused by relying on scalar edge weights in conventional graphs. Additionally, they proposed a Community Selection technique to reduce the computational overhead of spatial verification while maintaining high accuracy.

Robust multi-scale descriptors. The HOW model [32] improves image retrieval by efficiently combining global and local descriptors. It trains local descriptors indirectly using a contrastive loss on global optimization, which simplifies training by avoiding local correspondences. The model integrates local descriptors with the ASMK framework,

which enables robust and memory-efficient matching. With attention mechanisms prioritizing stronger features and multi-scale inference improving robustness, HOW achieved state-of-the-art performance on benchmarks like $\mathcal{ROxford}$ and $\mathcal{RParis}$ at the time. Also in 2020, Cao et al. [33] introduced the DELG (DEep Local and Global features) model, which unifies global and local image features into a single framework for efficient and accurate image retrieval. DELG employs a CNN to extract global features using GEM pooling, which effectively captures high-level, semantic information, and local features using an attention-based keypoint detector, optimized for discriminative regions. DELG also integrates a convolutional autoencoder for dimensionality reduction, streamlining feature compactness without additional post-processing. Hausler et al. [34] introduced Patch-NetVLAD, which combines local and global descriptors for robust place recognition. By deriving patch-level features from NetVLAD residuals and fusing multiple patch sizes, the approach enhances both spatial and appearance consistency across images. This multi-scale fusion mechanism enables the system to handle changes in viewpoint and illumination effectively. Yang et al. [35] introduced a single-stage Deep Orthogonal Local and Global (DOLG) method that integrates local and global features into a compact descriptor using an orthogonal fusion mechanism. The framework employs a global branch based on ResNet, enhanced with GeM pooling, and a local branch with multi-atrous convolution and self-attention to capture spatially discriminative features. The orthogonal fusion module extracts complementary local components orthogonal to the global features, ensuring efficient feature integration while reducing redundancy. Unlike traditional two-stage retrieval pipelines, DOLG bypasses the need for re-ranking and reduces error accumulation. Meanwhile, Wu et al. [36] proposed a novel Tokenizer module that aggregates deep local features into a compact set of visual tokens, each representing a specific visual pattern. This allows the model to focus on semantically meaningful regions while reducing the impact of background noise. The tokens are further enhanced using self-attention and cross-attention mechanisms to capture relationships across the image. This tokenization enables a highly compact global representation with strong regional matching capability. In 2022, Weinzaepfel et al. [37] introduced the FIRE model, which uses an iterative attention module that refines feature maps to focus on semantically meaningful and discriminative regions. Its novel features are called super-features. By using a contrastive loss combined with a decorrelation loss to diversify the features, this method balances the strengths of both local and global features and addresses the redundancy and mismatch issues commonly seen between local features and global image-level losses used in other retrieval systems. Similarly, Shao et al. [38] proposed SuperGlobal, a method using only global features for image retrieval and re-ranking, bypassing computationally expensive local feature matching. They optimized global feature extraction by enhancing GeM with Regional-GeM and Scale-GeM to effectively capture regional and multi-scale information. In the re-ranking stage, SuperGlobal updates global features dynamically based on the query and top-ranked images, significantly improving retrieval performance while being memory and compute-efficient.

Transformer-based architectures. Tan et al. [39] proposed Re-Ranking Transformers (RRTs), which combine global and local descriptors to efficiently re-rank top-matching images in image retrieval tasks. RRTs use a transformer to predict image pair similarity in a single pass, eliminating the need for computationally expensive geometric verification. This lightweight method outperformed traditional re-ranking techniques with fewer resources. Additionally, Wei et al. [40] introduced Visual Overlap Prediction (VOP), a novel method for place recognition focused on predicting the visual overlap between query and database images using patch-level embeddings. Their system leverages a Vision Transformer (ViT) backbone to establish patch-to-patch correspondences, moving away from traditional global similarity measures. VOP's use of a voting mechanism for

overlap prediction proved especially effective in handling occlusions and partial overlaps, leading to improvements in localization accuracy in large-scale indoor and outdoor datasets.

Enhancing nighttime retrieval. Mohwald et al. [41] proposed Dark Side Augmentation to target the challenge of image retrieval in difficult nighttime scenarios. They used a Generative Adversarial Network (GAN)-based synthetic-image generator to synthesize night images from daytime data. The approach significantly enhanced retrieval performance on challenging nighttime benchmarks while maintaining strong results on standard datasets.

2.2. Precise 2D-2D Localization

Precise 2D-2D localization focuses on obtaining an accurate 6-DoF camera pose by refining the initial result from the approximate 2D-2D localization. This refinement process, known as relative camera pose estimation, calculates the query image's pose relative to one or more retrieved database images. Since the poses of these database images are already known, the precise pose of the query image is determined through geometric transformations.

Precise 6-DoF monocular camera localization has traditionally been addressed using 2D-to-3D feature matching-based techniques, while using 2D-2D techniques for precise localization received limited research focus until the advent of deep learning. This is because conventional relative pose estimation methods depend on a sufficient number of matched key points between the query image and retrieved database images to accurately estimate the query's pose. Achieving this requires a significant overlap between the query image and the retrieved database images. However, in real-world scenarios, creating an image database for a large 3D environment with such dense coverage may not always be practical.

Classical geometry method. An early system for precise 2D-2D localization in urban environments was introduced by Zhang and Kosecka [42]. Their approach combined SIFT feature-based matching with a voting mechanism to identify candidate views, followed by robust motion estimation to determine the accurate camera pose. The motion estimation used either the essential matrix or a homography, depending on whether the points are in general position or lie on a plane. They also employed a RANSAC-based method to deal with the high percentage of outliers caused by repetitive structures in urban environments. However, in cases where reference views are sparsely distributed with minimal overlap, making reliable motion estimation difficult, the system resorts to interpolation for rough location estimates.

Deep learning-based approaches. With the rise of deep learning, CNN-extracted global features have proven effective in handling large viewpoint variations, leading to increased attention on precise 2D-2D localization. For instance, Melekhov et al. [43] introduced a CNN for relative camera pose estimation, which marked a shift towards deep learning in 2D-to-2D localization. Their Siamese network architecture took pairs of RGB images as inputs and directly estimated the relative rotation and translation between the cameras. This approach outperformed traditional feature-based methods (e.g., SURF and ORB) in handling large viewpoint variations. Laskar et al. [44] introduced Relative PN, a deep learning-based approach for precise 2D-to-2D localization. Their method utilizes a dual-branch Siamese CNN, pre-trained on the ImageNet dataset [45], to perform both approximate 2D-2D localization and relative pose estimation and then uses a RANSAC-based approach to triangulate the absolute camera position from the obtained relative poses. A significant advantage of this approach is that the CNN is trained on one dataset but can be applied to various environments, which enhances scalability and generalization. In addition, Balntas et al. [46] further advanced the field by introducing RelocNet,

which incorporated continuous metric learning for feature representation. Their method leverages the concept of camera frustum overlap to optimize the feature embedding network, ensuring that feature descriptors capture changes in camera pose. A pose regressor then infers finer relative poses between the query and nearest-neighbor images, using a geometric loss. This approach improved both retrieval accuracy and fine pose estimation, leading to more precise localization compared to earlier methods, especially in large-scale environments.

A lightweight architecture. Saha et al. [47] introduced the novel concept of anchor points for precise 2D-2D localization. AnchorNet, inspired by human landmark-based navigation, simultaneously predicts the most relevant anchor point in the scene and estimates the camera's relative offset from it. This multi-task learning approach uses CNNs to both classify the anchor point and regress the offsets. By focusing on anchor points, this method reduced the memory and runtime requirements compared to the multi-stage methods. However, its performance can be influenced by the visibility of anchor points.

Multi-stage pose refinement. Ding et al. [48] proposed CamNet, a triple-step coarse-to-fine localization framework based on a Siamese architecture that significantly improved the precision of 2D-2D localization. By separating the tasks of image-based coarse retrieval, pose-based fine retrieval, and precise relative pose regression into distinct modules, CamNet reduces task interference and refines pose estimates at each stage. Its use of novel loss functions, such as bilateral frustum loss and angle-based loss, further enhances retrieval and regression accuracy, particularly in scenes with large viewpoint variations.

Targeting structured environments. Li et al. [49] introduced a novel method for line-based absolute and relative pose estimation, specifically targeting structured environments. Their technique decoupled the rotation and translation components of the pose estimation process, proposing a two-step method for computing the rotation matrix. By leveraging the structural regularity of environments, such as parallelism and orthogonality, the method achieved highly precise pose estimates while being computationally efficient in environments with prominent structural features.

Combining classical and learning-based methods. Zhou et al. [50] presented a hybrid approach, EssNet, that combined classical feature-based methods with deep learning for relative pose estimation. They employed a Siamese-modified ResNet34 network with a fixed matching layer and introduced distinct matching layers to generate a score map for further regression and essential matrix optimization. Additionally, they integrated essential matrix estimation into a novel RANSAC-based framework to refine pose estimates. While the feature extraction was based on deep learning, the essential matrix regression maintained the reliability of classical methods.

Tackling a wide baseline. Chen et al. [51] tackled the challenge of wide-baseline relative camera pose estimation by introducing DirectionNet, which estimates a discrete distribution over the relative pose space instead of regressing poses directly. This method factorized relative camera poses into sets of 3D direction vectors and predicted them through CNNs, significantly improving pose estimation performance in low-overlap, large-motion scenarios.

3. 2D-3D Feature Matching-Based Localization

2D-3D feature matching-based localization (2D-3D localization), or structure-based localization, is the mainstream approach in monocular camera localization within known environments for precise 6DOF pose estimation. It relies on identifying and matching distinctive visual features between images and a pre-constructed 3D model of the environment. An illustration of 2D-3D feature matching is shown in Figure 4. While 2D-2D methods provide efficient but coarse localization, 2D-3D methods generally deliver 6-DoF pose estimation with significantly higher accuracy. However, the use of 3D models often comes

with increased computational demands. Based on the scale involved, 2D-3D localization methods can typically be divided into two sub-categories: direct 2D-3D localization and hierarchical 2D-3D localization.

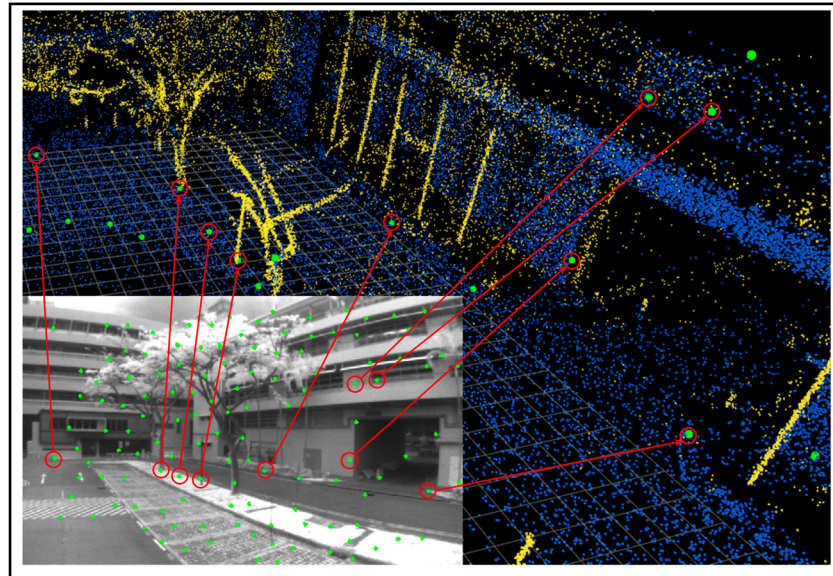


Figure 4. An example of matched features in 2D-3D matching. (Extracted features of the 2D image and 3D map are shown in green dots, while the correspondences between them are connected in red lines).

Direct 2D-3D localization involves matching the 2D image features directly to 3D scene features to establish correspondences and compute the camera pose. This approach relies heavily on the quality of feature detection and description to find precise correspondences between the query image and the 3D model, and it can achieve optimal localization in small- and medium-scale environments. On the other hand, hierarchical matching is an extension of direct matching that uses a hierarchical approach to first obtain rough camera pose estimations using 2D-2D approaches or sensors, such as GPS or a magnetic compass, and then uses the results to infer 2D-3D correspondences. This approach is efficient for large-scale environments, narrowing down the search space; however, it comes at a sacrifice in localization accuracy since the 3D map reduction can sometimes lead to a loss of necessary information. A flowchart illustrating the general workflow of these methods is shown in Figure 5.

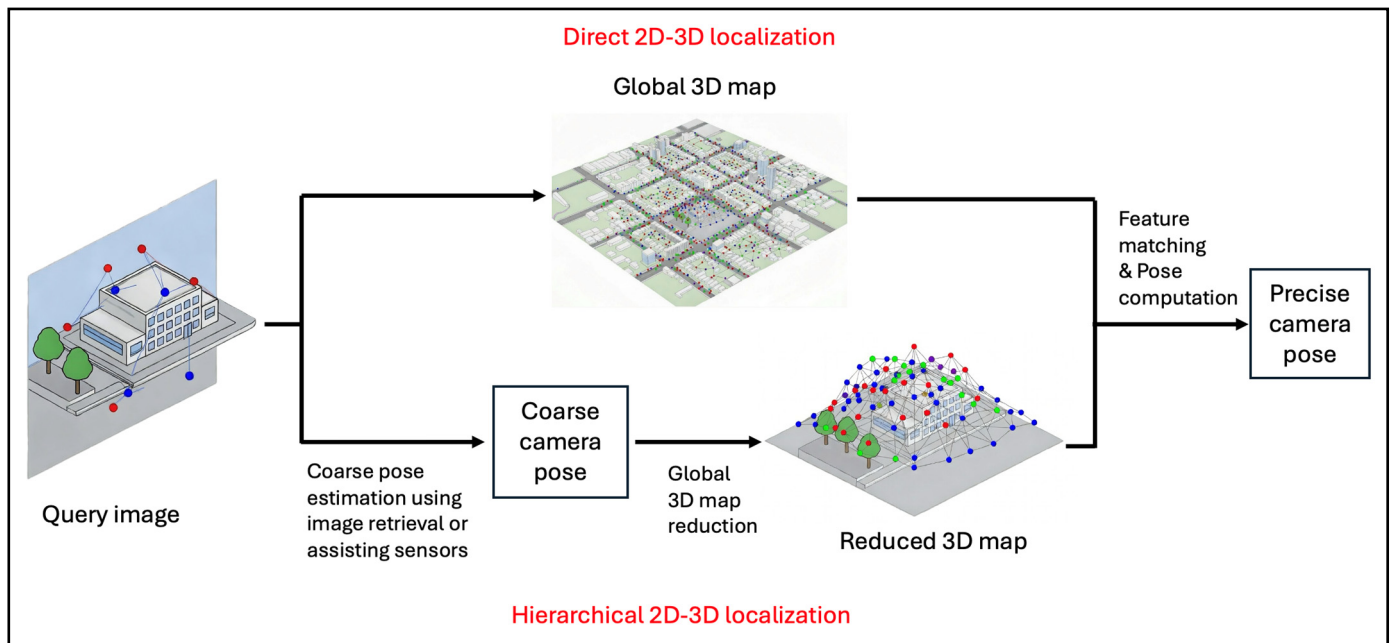


Figure 5. Overview of 2D-3D feature matching-based localization methods. (The top branch belongs to direct 2D-3D localization, while the bottom branch belongs to hierarchical 2D-3D localization.)

In the following, we review the methods for constructing 3D prior maps; techniques for direct matching and hierarchical matching in feature matching-based localization; and discuss the strengths, limitations, and applications of these techniques.

3.1. Creating a 3D Prior Map

An essential component of 2D-3D-matching-based localization is the construction of a 3D prior map of the environment. This map serves as a reference model against which features from new query images are matched to determine the 6-DoF camera pose. The 3D map can be constructed using various methods, including SfM, LiDAR-based scanning, and other point cloud generation methods.

Structure-from-Motion (SfM) [52]: SfM dominated the early stage of 2D-3D localization. It utilizes a collection of images captured from multiple viewpoints to reconstruct the three-dimensional geometry of the environment. Feature detection and matching algorithms identify correspondences between images and enable the triangulation of 3D points in space, resulting in a 3D point cloud that represents the structural layout of the environment. Each 3D point is associated with visual descriptors extracted from the images, creating a rich database that links visual appearance to spatial geometry.

LiDAR-based scanning: Recent studies have largely employed Light Detection and Ranging (LiDAR), which uses laser pulses to measure precise distances to objects. It can directly generate highly accurate and dense point clouds that capture the exact geometry of the environment, regardless of lighting conditions. LiDAR is particularly effective for large-scale outdoor environments and complex indoor spaces. The point cloud data from LiDAR can be combined with visual information by mapping textures or associating visual features, thereby enhancing feature matching techniques.

Beyond the widely used SfM and LiDAR, other technologies contribute to 3D map construction:

- **RGB-D sensors:** Devices like Microsoft Kinect and Intel RealSense capture both color and depth (i.e., RGB-D) information simultaneously. These sensors can generate real-

time 3D maps and are especially useful for indoor environments where lighting conditions are controlled.

- **Innovative Point Cloud Generation:** In response to the rapid advancement of sensor technologies, novel point cloud data types have emerged. These include multisource fusion point clouds [53] and Interferometric Synthetic Aperture Radar (InSAR) point clouds [54], which have shown exceptional value and distinct advantages in recent research [55–57].
- **Synthetic Data and Virtual Reconstructions:** Architectural models or Building Information Modeling (BIM) data can be used to generate virtual 3D environments [58–60]. These synthetic maps are beneficial for environments where physical data collection is impractical.

3.2. Direct 2D-3D Localization Methods

Direct 2D-3D localization techniques aim to directly establish correspondences between 2D features detected in query images and 3D features in a pre-constructed map of the environment. The core principle of direct matching involves three main steps: first, extracting distinctive features from the query images. These features are described using descriptors that capture the local appearance around each feature point. Second, matching the image feature descriptors to those associated with the 3D points in the map. Third, computing the camera pose from the established 2D-3D correspondences using a Perspective-n-Point (PnP) solver [61–65] inside a random sample consensus (RANSAC) loop [66–70].

3.2.1. SfM-Based Methods

The foundation of direct 2D-3D localization is rooted in Structure-from-Motion (SfM) research [71], which enables 3D points to acquire 2D visual features by generating 3D point clouds from 2D images.

Early foundation. The main bottleneck in early direct 2D-3D localization was the heavy computations and memory costs associated with retrieving features from large 3D models. Arth et al. [72] introduced an approach suitable for mobile devices by utilizing a sparse 3D reconstruction divided into potentially visible sets (PVSs). By precomputing these PVSs and loading only relevant portions of the model, the method enables efficient memory usage, making 2D-3D localization feasible on hardware with limited resources.

Optimizing feature matching. To target such issues more effectively, in the early 2010s, a focus on optimizing the matching process emerged. Li et al. [73] introduced a prioritized feature matching technique to improve efficiency. Instead of matching 2D features from a query image to all 3D points in the model, their approach prioritizes the most reliable 3D points based on their visibility across multiple views, thus efficiently reducing the number of comparisons needed and speeding up the localization process without sacrificing accuracy. Sattler et al. [74] expanded this work by introducing visual vocabulary quantization and a prioritized correspondence search, allowing for efficient localization in large-scale urban environments. This method laid the foundation for handling massive datasets by narrowing down the search space using visual words. Sattler et al. [75] then introduced active correspondence search, which combines 2D-to-3D and 3D-to-2D matching strategies to enhance the retrieval of 2D-3D correspondences. By guiding 3D-to-2D searches using initial 2D-to-3D matches, the nearby 3D points can also be considered during the matching process, significantly increasing the number of high-quality matches. Subsequently, Paudel et al. [76] addressed challenges in direct 2D-3D matching, such as unknown 2D-3D correspondences, outliers, and occlusions, by incorporating a constrained nonlinear optimization framework. This framework minimizes projection errors while preserving relationships between image pairs. The method also introduces a scale

histogram to resolve scale ambiguities and uses an M-estimator to reduce the effect of data inaccuracies.

Efficient memory and faster speed. Sattler et al. [77] improved memory efficiency and localization accuracy by introducing hyperpoints and fine visual vocabularies. Instead of directly matching descriptors, they proposed quantizing 3D descriptors into a fine visual vocabulary and representing each 3D point using multiple visual word IDs. Thus, each 2D query feature can be associated with a set of hyperpoints. This approach defers disambiguation during voting and preserves more information for matching, ensuring local uniqueness among correspondences. Also in 2015, Feng et al. [78] proposed using binary descriptors like BRISK and FRIF to speed up feature extraction. To counter the slower matching performance of binary features, they developed a supervised indexing approach using randomized trees. This approach brought significant speedups in matching performance while maintaining high accuracy. Sattler et al. [79] further improved the efficiency of large-scale 2D-3D localization by proposing an Active Search (AS) method, which prioritizes features likely to yield 2D-to-3D matches and terminates correspondence searches once sufficient matches are found. They also integrated 3D-to-2D matching to recover matches lost due to quantization, which leverages co-visibility information from the SfM process.

Handling large-scale environments. Liu et al. [80] focused on improving global 2D-to-3D matching accuracy in large-scale environments. They proposed a Markov network, which captures both the global contextual information and the co-visibility relationships between 3D points. By considering global compatibility between all matching pairs, the system can effectively resolve matching ambiguities in large-scale environments with repetitive structures.

Revisiting visual vocabularies. Song et al. [81] revisited the issue of quantization artifacts in visual vocabularies, which can cause lost matches and reduced localization accuracy. They proposed two mechanisms to recover lost matches: visibility-based recalling, which leverages co-visibility information from the 3D model, and space-based recalling, which relies on the spatial layout of features in the query image. These mechanisms improve the inlier ratio and localization accuracy with minimal computational overhead.

3.2.2. 3D-Scanner-Based Methods

While SfM-based methods have been instrumental in visual localization, they come with significant limitations. Creating 3D models with SfM is computationally expensive, and it tends to struggle with texture-less regions—often producing low-quality reconstructions or missing these areas entirely. In contrast, modern 3D-scanners like LiDAR, Microsoft Kinect, Matterport, and Faro offer high-quality, dense point clouds without the need for SfM. These scanners can capture large areas much more efficiently and with greater detail, particularly in texture-less regions where SfM fails, providing more accurate and robust data for localization. The direct 2D-3D localization methods in recent years have shown a significant transition from SfM-based methods to 3D-scanner-based methods.

Transition to 3D-scanner-based methods. The earliest transition came from Nadeem et al. [82], who bypassed the need for SfM by using LiDAR scanners to generate dense point clouds and directly matching 2D image keypoints (e.g., SIFT) with 3D point cloud keypoints (e.g., 3D-SIFT or RIFT). This foundational work introduced Descriptor-Matcher, trained on pairs of 2D and 3D descriptors, and set the stage for future research by offering a method to handle dense and detailed 3D data more efficiently. Building on this work, Nadeem et al. [83] introduced Unconstrained Matching (UM), which automatically collects a large number of matched 3D and 2D points and employs a two-stage coarse-to-fine

Descriptor-Matcher. This innovation improved the precision and computational efficiency for descriptor matching. Feng et al. [84] introduced 2D3D-MatchNet, an end-to-end deep learning framework. This approach separately extracts 2D features using SIFT and 3D features using ISS and then jointly learns 2D and 3D descriptors using deep networks, significantly improving robustness in 2D-3D correspondences. The deep learning shift marked the next step in robust, automated matching. Following the dual-pipe approach, Pham et al. [85] introduced a dual auto-encoder neural network, LCD, designed to learn cross-domain descriptors for matching between 2D images and 3D point clouds. The system maps 2D and 3D inputs into a shared latent space and improves descriptor discrimination by using a triplet loss function. The method outperforms traditional descriptor methods in matching accuracy.

Graph network. In 2020, Yu et al. [86] introduced a novel Bipartite Graph Network (BGNet) for 2D-3D matching. By constructing a bipartite graph of 2D points and 3D points, BGNet predicts the likelihood of each 2D-3D correspondence being an inlier. In addition, the method applies a Hungarian pooling layer to ensure optimal one-to-one matches, especially for challenging conditions such as varying illumination or repetitive patterns.

Targeting structured environments. Yu et al. [87] proposed extracting 3D geometric lines from LiDAR maps and matching them with 2D line features detected from video frames. This method utilizes visual-inertial odometry (VIO) for coarse pose prediction and iteratively optimized 2D-3D matches to minimize projection errors. The proposed system offers a lightweight solution for robust localization in structured environments.

Scene-agnostic approach. Sarlin et al. [88] presented PixLoc, which separates scene geometry from network parameters, enabling generalization without retraining. The model employs a multi-scale feature extraction pipeline using a CNN and refines poses through differentiable Levenberg–Marquardt optimization. A novel uncertainty-based weighting mechanism learns to ignore dynamic objects and repetitive patterns, focusing on stable scene elements.

Cross-domain and fully end-to-end learning. Lai et al. [89] introduced 2D3D-MVPNet, a network designed to learn cross-domain feature descriptors for matching between 2D image patches and 3D point cloud volumes. The method incorporates an image-based point cloud encoder to fuse multi-view projections of 3D point clouds with raw point cloud data, ensuring that both geometry and color texture information are captured. To handle the unordered nature of multi-view inputs, the authors introduced a multi-feature fusion module that uses a symmetric function. In 2023, Kim et al. [90] introduced a fully end-to-end approach, EP2P-Loc, that directly matches 3D point clouds to 2D image pixels without requiring keypoint detection. By employing a coarse-to-fine strategy, the method first retrieves relevant point cloud submaps and then uses 2D patch classification and positional encoding to locate the exact 2D pixel coordinates of the 3D points. EP2P-Loc also integrates a differentiable PnP solver, which enables efficient end-to-end 6DOF pose estimation.

Advancing descriptor representations. In 2022, Zhou et al. [91] proposed GoMatch, which eliminates the reliance on visual descriptors. By leveraging only geometric information through bearing vectors, GoMatch sidesteps the need for storage-heavy visual descriptors, significantly reducing storage requirements. Additionally, the method employs a self- and cross-attention mechanism to enhance matching robustness, particularly against outliers. In 2024, Nguyen et al. [92] introduced FUSELOC, which fuses global and local descriptors to enhance direct 2D-3D matching for visual localization. Using a weighted average operator, FUSELOC rearranges the local descriptor space to reduce search space ambiguity. This ensures that geographically distant descriptors are less likely

to compete with relevant ones. This fusion of descriptors significantly improves accuracy over local-only methods.

3.3. Hierarchical 2D-3D Localization Methods

Instead of relying on a global 3D map for localization, hierarchical matching techniques aim to reduce computational costs by limiting the size of the 3D map used for feature matching, though this often comes at the expense of accuracy. Depending on whether image retrieval or a positioning sensor is used, this is often achieved in one of two ways: (1) Image-retrieval-assisted methods: This approach retrieves database images similar to the query image and then uses the top K retrieved images to reduce the number of SfM points involved in the matching process. (2) Sensor-assisted methods: On the other hand, sensor-assisted methods estimate a rough camera pose using positioning sensors, like an Inertial Measurement Unit (IMU), gravity, compass, GPS, or radio signals (such as WiFi and Bluetooth). This initial pose narrows down the relevant portion of the 3D model for matching, which is then refined using a direct 2D-3D localization method. While SfM-based 3D models typically use the former or a combination of both strategies, 3D-scanner-based 3D models, where the 3D points are not linked to any 2D visual features, rely solely on the sensor-assisted approach.

Early image-retrieval-assisted methods. Irschara et al. [93] first introduced a fast image-retrieval-assisted 2D-3D localization method using SfM point clouds. The approach utilizes a vocabulary tree for efficient retrieval of relevant 3D point fragments directly from the point cloud, rather than searching an entire image database. Once these candidate points are identified, 2D-3D matching is performed between the query image features and the retrieved 3D point fragments. In addition, the authors proposed a compressed 3D scene representation to reduce memory usage while maintaining recognition accuracy. Despite the improvements in localization speed and efficiency offered by this hierarchical approach, direct 2D-3D methods remained dominant at the time due to their superior localization accuracy. To bridge this gap, Sattler et al. [94] revisited image-retrieval-assisted 2D-3D localization and analyzed the performance gap between hierarchical and direct matching methods. They introduced selective voting schemes and Hamming embeddings of feature descriptors to reduce false positives in hierarchical methods. This approach closes the performance gap between the hierarchical and direct matching while further improving the efficiency of localization.

Reducing 3D model size. Cao and Snavely [95] focused on reducing the size of SfM-based 3D models. They introduced a minimal scene description method focused on visibility coverage and feature distinctiveness. Their probabilistic framework employs a K-cover algorithm to select a minimal subset of 3D points while ensuring the distinctiveness of feature descriptors for robust localization. This approach significantly reduces memory consumption while maintaining good recognition performance. Additionally, Sattler et al. [96] questioned the necessity of using large-scale 3D models in direct 2D-3D localizations. Their system first retrieves relevant images based on a query and then constructs local 3D models around the retrieved images. The combination of image retrieval with local reconstructions provides pose estimates that are as accurate as structure-based methods while significantly reducing memory and computational costs. This hybrid approach suggested that large 3D models might not always be necessary for accurate visual localization, especially in large-scale urban environments.

Hybridization of direct and hierarchical methods. In 2018, Camposeco et al. [97] introduced a hybridization of direct and hierarchical 2D-3D approaches. They proposed a novel RANSAC-based framework that automatically selects the most appropriate solver for camera pose estimation. Depending on the quality of the 2D-3D and 2D-2D matches during each iteration, the system estimates the likelihood that a given solver will succeed,

allowing the system to flexibly switch between direct 2D-3D, image-retrieval-assisted 2D-3D, and hybrid solvers that use a combination of both types of correspondences.

Transition to deep learning. Taira et al. [98] leveraged deep visual descriptors for efficient localization. They proposed Inloc, which addresses large-scale indoor visual localization using a dense matching and view synthesis pipeline. Inloc consists of three key steps: (1) candidate pose retrieval using NetVLAD to retrieve top database images based on dense feature aggregation; (2) pose estimation via coarse-to-fine dense CNN feature matching with P3P-RANSAC for accurate camera pose estimation; and (3) pose verification using view synthesis, comparing the query image with a virtual view generated from the estimated pose. This approach addresses indoor-specific challenges such as textureless surfaces, repetitive structures, and cluttered scenes. In addition, Sarlin et al. [99] introduced an approach that emphasizes a coarse-to-fine global-to-local search strategy. It first retrieves candidate locations using NetVLAD global image descriptors to narrow down the search space. These retrieved keyframes are clustered based on shared 3D points (co-visibility clustering), limiting the scope for local 2D-3D matching. Only points within these clusters are used for precise pose estimation, greatly reducing computational demands. Building on this work, Sarlin et al. [100] introduced HF-Net, which integrates global and local feature extraction into a single neural network. A coarse global retrieval step first identifies potential locations, followed by detailed local matching of 2D features to 3D points for precise pose estimation. The shared computation across both steps enhances robustness and reduces runtime, making it scalable for large environments with significant appearance changes. Furthermore, Dusmanu et al. [101] proposed D2-Net, a unified framework where a single CNN detects and describes local features by extracting dense features and detecting keypoints as maxima of feature maps. Unlike traditional methods that separate detection and description, D2-Net ensures better consistency between the two. It excels under challenging conditions like illumination changes and view-point variations. In 2020, Germain et al. [102] proposed S2DNet, a sparse-to-dense matching framework. S2DNet leverages dense feature maps and treats the correspondence task as a supervised classification problem. It extracts multi-level features using a VGG-16 backbone and computes correspondence maps at multiple levels, aggregated via bilinear upsampling. Additionally, Yang et al. [103] proposed a unified framework, UR2KiD, for image retrieval and local matching, trained without dense ground truth or SfM models. Using a ResNet-based architecture, it combines global descriptors via GeM pooling and local keypoint detection through a Group-Concept Detect-and-Describe module, which groups feature channels and thresholds based on L2 norms. Joint optimization across retrieval, matching, and self-distillation tasks enables robust performance under challenging conditions like scale and illumination changes. Also in 2020, Shi et al. [104] introduced Dense Semantic Localization (DSL_{Loc}), which integrates handcrafted and learned features with a dense semantic 3D map for robust localization. The dense semantic map enables reliable 2D-3D matching by ensuring semantic consistency between query images and 3D points. This consistency score is incorporated as a soft constraint in a weighted RANSAC-based PnP solver.

Hybridization of Image retrieval and positioning sensors. Humenberger et al. [105] introduced the Kapture framework, which integrates multiple sensors such as GPS, IMU, and WiFi into the image-retrieval-assisted pipeline. The process starts with image retrieval to reduce the search space, followed by sensor data to further constrain the search. GPS narrows down the geographic area for retrieval, and IMU data helps improve orientation estimates during pose estimation. This combination of sensor data with image retrieval and 2D-3D matching enables robust and efficient localization, especially in dynamic and large-scale environments.

Integrating MEMS with direct 2D-3D localization. Shu et al. [106] introduced a novel hybrid approach that integrates MEMS (Micro-Electro-Mechanical Systems) sensors with 2D-3D localization techniques to address the computational and memory challenges of indoor localization on mobile devices. The MEMS-based module uses Pedestrian Dead Reckoning (PDR) and the Attitude and Heading Reference System (AHRS) to provide initial rough 6-DoF pose estimates, greatly reducing the search area of the 3D model. Subsequently, direct 2D-3D localization is used to match 2D features to 3D points while a geometric pose error propagation model is applied to filter outliers during the RANSAC-based pose estimation. Finally, a dynamic reliability strategy fuses the MEMS and 2D-3D results to continuously refine the 6-DoF pose.

Multi-sensor fusion. Yan et al. [107] introduced the SensLoc framework, which integrates multiple mobile sensors—GPS, compass, and IMU sensors—with 3D-scanner-based 2D-3D localization. The framework uses GPS and compass data to constrain the search space of the 3D model, followed by a coarse-to-fine transformer-based network to perform direct 2D-3D matching. Lastly, they adopted a gravity-guided PnP RANSAC, which utilizes the IMU pitch and roll angles to filter out incorrect pose hypotheses, significantly enhancing both the efficiency and accuracy of the final pose estimation.

4. Regression-Based Localization

This section reviews various regression-based localization methods. These methods use deep neural networks to predict spatial relationships from input images directly, bypassing the explicit feature matching in 2D-2D and 2D-3D localization methods. Two main categories of regression-based localization methods are discussed in this section: absolute pose regression (APR) and scene coordinate regression (SCR). APR methods directly predict the camera's absolute position and orientation from an input image using neural networks, whereas SCR methods predict a 3D scene coordinate for every pixel in an input image, followed by pose estimation using these coordinates. A flowchart illustrating the general workflow of these methods is shown in Figure 6.

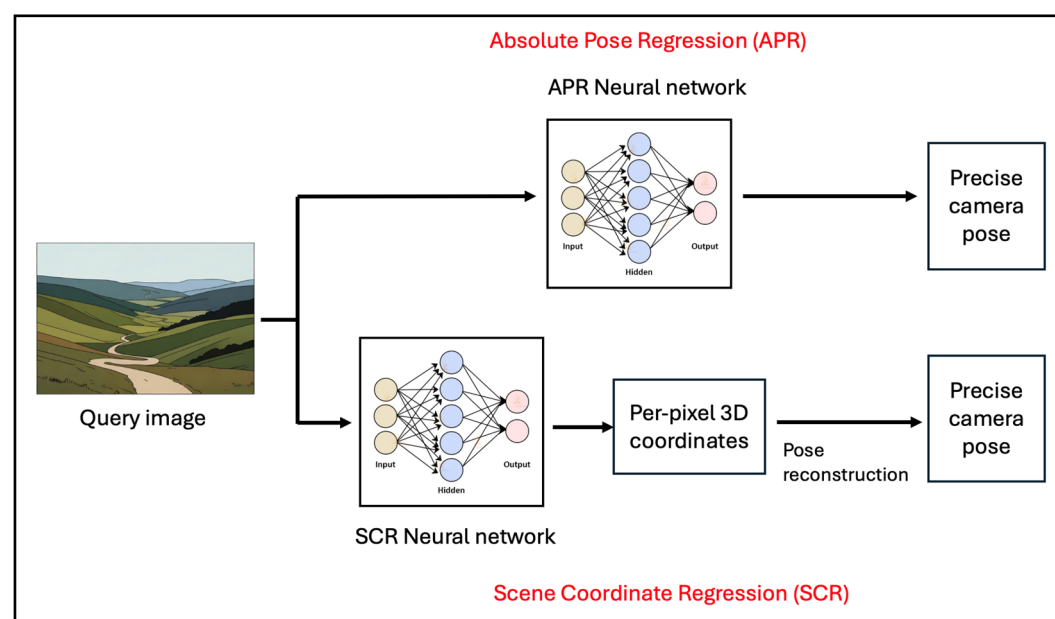


Figure 6. Overview of regression-based localization methods. (The top branch belongs to absolute pose regression, while the bottom branch belongs to scene coordinate regression.)

4.1. Absolute Pose Regression

Absolute pose regression (APR) leverages deep neural networks to directly predict a camera's 6-DoF pose from a single image. Unlike 2D-2D or 2D-3D matching-based techniques, which rely on correspondences between image features and pre-built maps, APR uses end-to-end learning to map an image directly to its absolute pose, making it highly appealing for scenarios where rapid pose estimation is crucial or where building detailed maps is challenging.

The pioneering PoseNet. Kendall et al. [108] introduced a pioneering APR method, PoseNet, that uses a CNN to directly regress the 6-DoF camera pose from a single RGB image. PoseNet innovatively avoids explicit feature matching or the construction of 3D maps. Instead, it modified an image classification CNN by replacing the softmax layers with fully connected layers to regress both the camera 3D location and orientation. In addition, it leverages transfer learning from large-scale image classification tasks and uses SfM for automatic labeling of training data. PoseNet achieved reasonable accuracy and operated in real time. However, its localization accuracy was not on par with the traditional methods relying on 2D-3D matching.

PoseNet Variants. To reduce the localization error of the PoseNet architecture, multiple subsequent improvements have been proposed, including incorporating a dense cropping technique, a probabilistic architecture, a new loss function, or Long Short-Term Memory (LSTM) layers. In addition to the standard PoseNet, Kendall et al. [108] introduced a dense cropping technique, referred to as Dense PoseNet, which improved PoseNet accuracy by averaging pose predictions from 128 uniformly spaced crops of the input image. The dense version of PoseNet, while computationally more intensive (increasing processing time from 5 ms to 95 ms), significantly enhanced its performance by incorporating more spatial information during inference, leading to more precise pose predictions. Building on PoseNet, Kendall et al. [109] introduced a probabilistic framework, Bayesian PoseNet, using Bayesian convolutional neural networks to model uncertainty in pose predictions. It employs Monte Carlo dropout at inference time, which provides uncertainty estimates alongside pose predictions, improving robustness to ambiguous scenes and out-of-distribution inputs. Bayesian PoseNet shows enhanced localization accuracy while maintaining real-time performance (under 6 ms per image) and requiring minimal memory. In the follow-up of PoseNet, Kendall et al. [110] introduced GeoPoseNet, which incorporates geometric loss functions to improve the performance of PoseNet. They emphasized that the original PoseNet employs a naive loss function that does not consider the relationship between position and orientation. This new version optimizes the regression process by incorporating scene reprojection error and automatically learning the optimal weighting between position and orientation, significantly improving localization accuracy. Subsequently, Walch et al. [111] expanded upon the PoseNet framework by incorporating LSTM units into the architecture. The LSTMs reduce the dimensionality of the high-dimensional outputs from the fully connected layers of a PoseNet. They act as a form of structured regularization that allows the network to focus on the most important feature correlations for pose regression. The LSTM PoseNet improved localization accuracy by 32% and 37% with respect to the PoseNet and Bayesian PoseNet.

HourGlass-Pose. Besides the PoseNet architecture, Melekhov et al. [112] proposed an encoder–decoder CNN with an hourglass-shaped architecture for camera pose estimation. The encoder captured global scene context, while the decoder used up-convolution layers and skip connections to recover fine-grained spatial information, improving localization accuracy.

BranchNet. Wu et al. [113] introduced BranchNet along with several key advancements. They introduced Euler6, a new orientation representation that addresses issues with quaternion representation in CNN-based pose regression. They also introduced pose

synthesis, a novel data augmentation technique to address pose sparsity by generating new poses through image patch translations. Finally, they proposed BranchNet, a multi-task CNN that separately regressed orientation and translation in distinct branches, significantly reducing localization errors compared to Bayesian PoseNet.

SVS-Pose. Naseer and Burgard [114] proposed SVS-Pose. It uses a modified VGG16 architecture with separate fully connected layers to predict camera position and orientation. To improve performance, the authors employed 3D data augmentation, generating synthetic viewpoints from single images to expand the training pose coverage.

MapNet. Brahmbhatt et al. [115] introduced MapNet. MapNet incorporates geometric constraints from visual odometry (VO), GPS, and IMU sensors into its training process. The model regresses camera poses from images while also enforcing relative pose constraints between image pairs to improve global consistency. Additionally, the authors propose MapNet+, which uses unlabeled video data to update the learned maps in a self-supervised manner, and MapNet+PGO, which integrates pose graph optimization (PGO) to refine pose predictions during inference. This framework significantly improves localization accuracy by fusing data from multiple sensory inputs and continuously refining maps as more data becomes available.

Self-attention mechanism. In addition, Wang et al. [116] introduced AtLoc, an attention-guided camera localization framework that improves the robustness of single-image pose estimation. The method utilizes a self-attention mechanism to focus on geometrically stable and informative regions in the scene, such as building structures, while ignoring dynamic elements like moving vehicles. The combination of attention with a CNN-based encoder led to superior performance, even in complex outdoor scenes with high variability.

Improved probabilistic architecture. Building on the concept of the Bayesian PoseNet, Cai et al. [117] proposed GPoseNet, which combines CNNs and Gaussian Process Regression (GPR) to model uncertainty in 6-DoF camera localization. The system leverages CNNs to extract features from RGB images and uses sparse variational inference GPs for translation and rotation, providing a distribution over camera poses in a single inference step, which significantly improves computational efficiency. In addition, the objective function uses Kullback–Leibler divergence to balance the learning of both translation and rotation, which reduces sensitivity to hyperparameters.

Multi-scene APRs. Chidlovskii and Sadek [118] introduced a multi-scene adaptation approach, APANet, using adversarial learning to address the challenge where models trained on one scene struggle to generalize to unseen scenes. The method trains an encoder to produce scene-invariant image representations, which allows the model to generalize across different environments. A scene discriminator is used in a minimax optimization framework to distinguish between source and target scene features, while the encoder was trained to deceive the discriminator, thus achieving better model transfer. The approach also incorporated self-supervised learning by predicting image rotations, which enhances the model's scene understanding. Shavit and Ferens [119] challenged the need for scene-specific encoders in APR by proposing IRPNet, a branching multilayer perceptron (MLP) that uses pre-trained image retrieval (IR) models to generate generic visual encodings for pose estimation. Instead of fine-tuning for each scene, IRPNet leverages IR-based global descriptors to regress the camera's position and orientation. This approach significantly reduces training time and memory usage while achieving competitive results across datasets. The results show that scene-agnostic encodings can effectively replace scene-specific models. Following the multi-scene adaptation concept, Blanton et al. [120] introduced Multi-Scene PoseNet (MSPN), which handles pose regression across multiple scenes using a shared feature extraction network followed by a scene-specific layer that adapts the model's predictions based on the scene. The model also uses a scene-prediction

component to index scene-specific weights, which reduces memory usage while maintaining strong performance across diverse datasets. MSPN demonstrated improved generalization over single-scene models and provided better initialization for fine-tuning on new scenes. Subsequently, Shavit et al. [121] proposed a transformer-based architecture, MS-Transformer, for multi-scene absolute pose regression. The model uses separate transformers to process position and orientation features extracted from a convolutional backbone. These transformers attend to informative features using self-attention and produce scene-specific pose predictions.

Enhancing accuracy using auto-encoders. Shavit and Keller [122] introduced Camera Pose Auto-Encoders (PAEs) to enhance their previous MS-Transformer by leveraging prior pose information efficiently. PAEs are trained using a teacher–student approach, where the teacher APR’s outputs are used to train the auto-encoder to generate latent pose representations. These encodings, robust to appearance changes, are used to refine camera pose estimates during test time, significantly enhancing position accuracy without needing additional images.

Exploiting motion constraints. Clark et al. [123] introduced VidLoc, a deep spatio-temporal model for 6-DoF camera localization using video clips. It combines a CNN for image feature extraction with a bidirectional LSTM to capture temporal dependencies between frames. By leveraging both past and future frame information, VidLoc improves the accuracy and smoothness of pose estimates. The model is trained with a loss function that balances translation and rotation errors, and it incorporates uncertainty estimation using mixture density networks. VidLoc demonstrated significant accuracy improvements over single-image methods like PoseNet. Subsequently, Valada et al. [124] proposed an auxiliary learning model, VLocNet, which incorporates relative pose estimation as an auxiliary task to constrain the search space and improve consistency in pose regressions from consecutive monocular images. The model employs a three-stream architecture: one stream for global pose regression and a Siamese-type double-stream for relative pose estimation. Shared parameters across tasks allow VLocNet to capture inter-task correlations, while a Geometric Consistency Loss enforces alignment between global and relative poses. Building on this work, Radwan et al. [125] proposed VLocNet++. The model improves VLocNet by employing a novel adaptive weighted fusion layer that dynamically integrates motion-specific temporal features and semantic information, improving pose regression accuracy. Also, it incorporates a self-supervised warping technique that leverages relative motion to aggregate scene context, enhancing segmentation consistency and accelerating training. In 2019, Bui et al. [126] proposed AdPR. They extended the idea of using temporal sequences for camera localization by leveraging local information from image sequences to improve global pose estimation accuracy. First, a convolutional LSTM-based visual odometry model predicts relative poses from consecutive frames. Then, in a content-augmentation step, co-visible features from local maps are integrated using soft attention to refine the global pose estimates. Finally, a motion-based refinement step follows, using a pose graph to optimize global poses based on VO-derived relative motions. This two-step process significantly improved accuracy in challenging scenarios with repetitive or low textures. In 2023, Wang et al. [127] proposed RobustLoc, which employs neural differential equations within a graph-based structure to enhance robustness against environmental perturbations like weather, lighting changes, and dynamic objects. The model uses a CNN to extract feature maps from multi-view images, followed by a neural diffusion block to diffuse information across neighboring frames, capturing both local and global context. A branched decoder with multi-level training then regresses the 6-DoF pose estimates.

Multi-Level Features (MLF). Xu et al. [128] developed a framework that integrates local and global features for enhanced robustness. Their methodology uses a multi-level

deformable network to extract local features that are sensitive to rotation and other transformations. These local features are combined with global features incorporating geometric constraints to improve pose estimation accuracy. In addition, they proposed a novel loss function that simultaneously minimizes the regression error of both relative and absolute pose estimates. This approach demonstrated superior accuracy in challenging conditions, particularly when dealing with significant viewpoint and illumination changes.

DFNet and DFNet+. Chen et al. [129] introduced a hybrid camera localization pipeline combining APR and direct feature matching for robust pose estimation. Its core includes two components: a Histogram-assisted NeRF for view synthesis, reducing the photometric domain gap by adapting image appearance based on luminance histograms, and DFNet, which extracts domain-invariant features using a contrastive loss, bridging the gap between synthetic and real images. Additionally, it employs Random View Synthesis for efficient synthetic data generation, improving pose regression accuracy with minimal supervision. Building on this work, Chen et al. [130] proposed DFNet+, which employs a novel Neural Feature Synthesizer (NeFeS) that predicts dense feature maps from synthesized views. During inference, NeFeS compares rendered and query features, backpropagating feature-metric errors to refine camera poses. Additionally, they introduced an Exposure-Adaptive Affine Color Transformation for handling illumination changes and a Feature Fusion Module for merging rendered features and RGB values.

4.2. Scene Coordinate Regression

Unlike APR methods, which directly regress camera pose, SCR methods predict 3D scene coordinates from 2D image pixels, followed by camera pose estimation using a PnP solver, similar to 2D-3D localization methods. Over the years, SCR methods have undergone a transition from early randomized regression forests to recent deep learning-based methods. Early SCR methods [131–133] heavily relied on regression forests using RGB-D images, which leverage depth data as a crucial feature for 3D scene coordinate prediction. The reliance on RGB-D data persisted until the introduction of deep learning-based approaches.

DSAC series. Brachmann et al. [134] first introduced an end-to-end trainable camera localization pipeline based on Differentiable Sample Consensus (DSAC). The method replaces RANSAC's non-differentiable hypothesis selection with a probabilistic selection mechanism inspired by reinforcement learning, enabling backpropagation through the entire pipeline. The system consists of two CNNs: one predicts 3D scene coordinates from RGB images, and the other scores pose hypotheses based on reprojection errors. Hypotheses are generated using the PnP algorithm, and the best pose is selected probabilistically. Subsequently, Brachmann et al. [135] introduced DSAC++, which replaces the learnable hypothesis-scoring CNN used in DSAC with a soft inlier count mechanism, significantly improving generalization and reducing overfitting. In 2019, Brachmann et al. [136] proposed Expert Sample Consensus (ESAC), which improves upon DSAC by addressing scalability and ambiguity in challenging environments through a Mixture of Experts (MoE) approach. Unlike DSAC, which uses a single neural network to predict 2D-3D correspondences, ESAC employs multiple specialized expert networks, with a gating network assigning probabilities to experts based on the input. Hypotheses from all experts are evaluated using DSAC's geometric consistency scoring, ensuring robustness against gating errors. In 2021, Brachmann et al. [137] further proposed DSAC* by replacing the two-stage initialization process of DSAC++ with a unified training objective. In addition, the authors updated the VGG-style architecture used in DSAC++ with a ResNet-based fully convolutional network. These enhancements reduce training time while enhancing accuracy.

HSCNet series. Li et al. [138] introduced an SCR framework, HSCNet, which combines classification and regression layers to predict 3D scene coordinates in a hierarchical coarse-to-fine manner. The network first predicts coarse location labels through discrete classification layers and progressively refines predictions using sub-region labels. A final regression layer estimates precise 3D coordinates relative to the predicted clusters. This design improves scalability, resolving visual ambiguities common in large environments. Subsequently, Wang et al. [139] introduced HSCNet++, which replaces convolutional layers with transformer-based modules for global context aggregation. Additionally, the system introduces a pseudo-labeling strategy for training with sparse 3D annotations, combined with an angle-based reprojection loss for robustness against noisy labels.

Incorporating probabilistic modeling. Rekavandi et al. [140] introduced B-Pose, a Bayesian CNN for camera localization that combines scene coordinate regression with uncertainty modeling. It predicts dense 3D scene coordinates using a ResNet backbone and refines pose estimates via Differentiable RANSAC and Monte Carlo sampling for uncertainty quantification.

NeuMap. Tang et al. [141] introduced NeuMap, which encodes scenes into latent codes and uses a transformer-based auto-decoder for SCR. NeuMap is designed to address the challenges of large-scale scenes by dividing them into smaller voxels, using latent codes to store scene information. Upon inference, query image features are matched to these codes through a cross-attention mechanism, yielding 2D-3D correspondences for PnP-based pose estimation. The system supports efficient fine-tuning for new scenes and small scene representation sizes.

Scene geometry reasoning. Chen et al. [142] introduced MaRepo, a framework combining pose regression with scene-specific geometric reasoning. Unlike traditional APR, MaRepo leverages a scene geometry network to predict 3D scene coordinates and a transformer-based pose regressor to estimate camera poses. Key innovations include dynamic positional encoding for camera intrinsics and spatial embeddings, and a re-attention mechanism for enhanced information flow. MaRepo enables rapid adaptation to new scenes.

Scene-Agnostic SCR. Revaud et al. [143] proposed SACReg, which generalizes across diverse scenes without the need for scene-specific fine-tuning. The method adopts Vision Transformers to encode both the query and database images, with database tokens augmented using sparse 2D-3D correspondences. A transformer decoder fuses this information and predicts dense 3D coordinates along with confidence scores for each pixel. To ensure generalization, the model uses a novel cosine-based coordinate encoding that supports arbitrary scene sizes, overcoming limitations of previous scene-specific SCR models.

Lightweight models. Brachmann et al. [144] proposed an Accelerated Coordinate Encoding (ACE) network that learns dense 2D-3D correspondences from posed images through contrastive pose supervision, which maximizes localization accuracy by comparing correct and incorrect pose estimates. A novel geometric consistency loss enforces spatial coherence between predicted scene coordinates. Subsequently, Lu et al. [145] proposed ACE++, which improves localization accuracy over ACE while maintaining rapid training. Building upon ACE, the method introduces two key innovations: (1) coordinate convolution layers, which add spatial awareness by encoding 2D positional information directly into feature maps, enhancing scene coordinate regression capability, and (2) a Group Convolutional Block Attention Module, which dynamically refines channel and spatial attention to focus on important features.

5. Datasets and Evaluation Metrics

In monocular camera localization, the availability of high-quality datasets and robust evaluation metrics is crucial for advancing research and benchmarking new methods. This

section provides a summary of the key datasets commonly used in this field and the evaluation metrics used to measure the localization accuracy.

5.1. Benchmark Datasets

Benchmark datasets for monocular camera localization cover a range of scenarios, from indoor environments with consistent lighting to complex outdoor scenes with a variety of conditions, including day–night, weather, and seasonal variations. These datasets can be divided into two main categories: place recognition datasets and 6-DoF pose estimation datasets. Since approximate 2D-2D localization techniques adopt the pose of the nearest neighbor as the localization result, these methods are primarily evaluated using place recognition (or image retrieval) datasets. In contrast, 6-DoF pose estimation datasets are largely used to evaluate precise 2D-2D localization, 2D-3D localization, and regression-based localization techniques.

5.1.1. Place Recognition Datasets

Place recognition and image retrieval datasets are widely used to evaluate approximate 2D-2D localization techniques. Below, we summarize the commonly used place recognition datasets in the context of monocular camera localization in known environments, with key details provided in Table 1.

University of Kentucky Benchmark (UKB) dataset [10]: The UKB dataset is a widely used object-focused image retrieval benchmark. It consists of 10,200 indoor images of 2550 unique objects from diverse categories, such as animals, plants, household objects, etc., with each object captured from four different viewpoints. Every image in the dataset is used as a query during evaluation, while accuracy is calculated from the top four retrieved results (precision@4). The dataset is particularly notable for its controlled setup and focus on object-level similarity.

Oxford Buildings dataset [11]: The Oxford Buildings dataset contains two sub-datasets: Oxford5K and Oxford105K. Oxford5K consists of 5062 high resolution (1024×768) images collected from Flickr, an online photo-sharing and hosting platform. These images feature various landmarks in Oxford, UK, and are annotated with 11 landmark classes. The dataset includes 55 query images, with ground truth relevance provided for evaluating retrieval performance. Due to its focus on urban landmarks, this dataset is particularly suited for assessing the robustness of localization algorithms in structured and repetitive architectural environments. On the other hand, Oxford105K extends the standard Oxford5K dataset by including an additional 99,782 distractor images from Flickr. The added distractor images pose challenges for retrieval algorithms in distinguishing relevant landmarks from irrelevant data, thereby enhancing their utility for large-scale localization evaluations.

Paris Buildings dataset [146]: Similar to the Oxford Buildings dataset, the Paris Buildings dataset provides robust benchmarks for evaluating algorithms in urban landmark recognition tasks. It contains two sub-datasets: Paris6K and Paris106K. Paris6K contains 6412 high-resolution (1024×768) images collected from Flickr, which are associated with various iconic landmarks in Paris, France, such as “Paris Eiffel Tower” or “Paris Triomphe”. The Paris106K dataset extends Paris6K by adding 100,000 distractor images, significantly increasing the complexity and scale of image retrieval tasks.

INRIA Holidays dataset [147]: The Holidays dataset is widely used to evaluate the robustness of place recognition and image retrieval methods. It comprises 1491 high-resolution vacation images organized into 500 groups, with each group representing a distinct scene or object. A single image from each group serves as the query, while the rest represent the correct retrieval results. This dataset is particularly notable for its diverse range of outdoor scene types (natural, man-made, water and fire effects, etc.).

24/7-Tokyo dataset [20]: The 24/7 Tokyo dataset was developed to evaluate place recognition techniques under significant illumination changes, including day, sunset, and night conditions. It consists of 1125 images captured at 125 locations across Tokyo, Japan, with images taken in three viewing directions and at three times of the day for each location. Ground truth locations are GPS-annotated with an accuracy of within 5 m. The dataset emphasizes substantial variations in lighting and structural elements, making it ideal for testing algorithms' robustness to extreme appearance changes.

Revisited Oxford (ROxford) and revisited Paris (RParis) datasets [148]: The revisited Oxford and revisited Paris datasets are updated benchmarks for the original Oxford5K and Paris6K datasets. They address issues in the original datasets, including annotation errors, dataset size, and difficulty level. In total, 15 challenging queries are added to each dataset, increasing the total to 70 queries. In addition, images are relabeled into four categories, easy, hard, unclear, and negative, based on viewpoint, occlusion, and illumination. Based on these difficulty labels, three difficulty levels are introduced for evaluations: easy (includes only easy images as positives), medium (includes both easy and hard images as positives), and hard (Includes only hard images as positives). In addition, a new distractor set of 1,001,001 challenging images, R1M, was introduced, which can be added to ROxford or RParis to significantly increase the scale and complexity of evaluations.

Google Landmarks dataset (GLD) [27]: The GLD is a large-scale dataset designed for evaluating image retrieval and landmark recognition tasks. It contains over 1 million images from 12,894 landmarks worldwide, along with an additional 100,000 query images. The dataset introduces challenges such as background clutter, occlusion, and diverse landmark types, including human-made and natural structures.

Google Landmarks dataset v2 (GLDv2) [149]: GLDv2 builds upon the original GLD with over 5 million images and 200,000 distinct instance labels, making it the largest instance recognition dataset to date. It includes a training set of 4.1 million labeled images, 762,000 index images, and a test set of 118,000 query images. GLDv2 introduces several challenges—a long-tailed class distribution, high intra-class variability, and a significant fraction of out-of-domain test images—to simulate real-world applications. The dataset emphasizes practical challenges, such as robustness to false positives and scalability in large-scale settings.

Table 1. Mainstream place recognition datasets.

Dataset Name	Scale	Scene	Challenging Factors	Image Number
UKB [10]	Small	Indoor	Scale, illumination, background clutter	10,200
Oxford5K [11]	Medium	Urban landmarks	Scale, occlusion, repetitive structures	5011 references, 55 queries
Oxford105K [11]	Large	Urban landmarks	Scale, occlusion, repetitive structures	104,793 references, 55 queries
Paris6K [146]	Medium	Urban landmarks	Scale, occlusion, repetitive structures	6245 references, 55 queries
Paris106K [146]	Large	Urban landmarks	Scale, occlusion, repetitive structures	106,031 references, 55 queries
INRIA Holidays [147]	Medium	Diverse urban and suburban scenes	Illumination, viewpoint, scene diversity	991 references, 500 queries
24/7-Tokyo [20]	Large	Urban streets	Day–night,	75,984 references, 1125 queries,

			season, dynamic urban conditions	597,744 synthetic views
$\mathcal{R}$ Oxford [148]	Medium	Urban landmarks	Scale, occlusion, illumination	4993 references, 70 queries
$\mathcal{R}$ Paris [148]	Medium	Urban landmarks	Scale, occlusion, illumination	6322 references, 70 queries
GLD [27]	Large	Global natural and human-made landmarks	Occlusion, clutter, scale, partial/multi-landmarks	1,060,709 references, 111,036 queries
GLDv2 [149]	Large	Global natural and man-made landmarks	Occlusion, clutter, scale, partial/multi-landmarks	4.1 M references, 118 K queries

5.1.2. 6-DoF Pose Estimation Datasets

In contrast to the place recognition datasets, 6-DoF pose estimation datasets are designed to evaluate localization techniques, where determining an exact 6-DoF camera pose is essential. The commonly used 6-DoF pose estimation datasets are summarized as follows, with key details provided in Table 2.

Vienna dataset [93]: The Vienna dataset is a preliminary dataset designed for image-based localization and urban scene reconstruction. It focuses on parts of Vienna, Austria, captured by car-mounted cameras with 1324 high-resolution reference and 266 query images. The dataset includes a 3D model with 1,123,028 3D points generated through SfM. Challenges include varied illuminations, weather, and seasonal changes.

Rome dataset [73]: The Rome dataset is a large-scale dataset designed for image-based localization. It is built from a collection of 16,179 geotagged images sourced from Flickr, covering various landmarks in Rome. Using SfM techniques, the dataset generates a 3D point cloud with 4,312,820 points, organized into 69 connected components, representing different landmarks or scenes.

Carnegie Mellon University (CMU) dataset [150]: The CMU dataset is a visual localization benchmark that emphasizes seasonal and environmental variability. A vehicle equipped with two cameras on the roof at 45-degree angles captures 256×192 -resolution images from urban, suburban, and park-like environments along an 8.8 km route in Pittsburgh, USA. Approximately 10,000–14,000 images are stored per full traversal. The dataset includes GPS-annotated images and topometric mapping. Challenges in the dataset include illumination changes, occlusions, and structural modifications.

Aachen dataset [94]: The Aachen dataset is designed to test image-based localization methods in challenging urban environments. It features 4479 high-resolution reference images, 369 query images, and a 3D model reconstructed using SfM. The images are taken in the old inner city of Aachen, Germany, with varying conditions such as different times of day, weather changes, and occlusions. This dataset emphasizes real-world complexities such as motion blur, lighting variations, and environmental changes.

7-Scenes dataset [131]: 7-Scenes is a widely used benchmark for camera pose estimation in indoor environments. It consists of RGB-D image sequences at a resolution of 640×480 pixels in seven distinct scenes, including chess, fire, heads, office, pumpkin, red kitchen, and stairs. Each scene features challenges like motion blur, texture-less surfaces, repeated structures, and varying lighting conditions. The dataset provides ground truth 6-DoF camera poses derived from KinectFusion-based tracking. The dataset is particularly valuable for testing algorithms in complex indoor settings.

Cambridge Landmarks dataset (CLD) [108]: The CLD is an outdoor visual localization benchmark captured around the University of Cambridge. It comprises five distinct urban scenes: King's College, Old Hospital, Shop Façade, St. Mary's Church, and Street. The dataset comprises images captured with a handheld smartphone camera at 640×480

resolution, with ground truth 6-DoF poses obtained through SfM. The dataset introduces significant challenges, such as urban clutter, dynamic elements like pedestrians and vehicles, and variations in lighting and weather.

RobotCar dataset [151]: The RobotCar dataset is a comprehensive benchmark for evaluating visual localization algorithms in dynamic urban streets. Captured in Oxford, UK, the dataset features over 100 traversals of a consistent 10 km route of driving data recorded across different seasons, times of day, weather conditions, and dynamic elements like traffic and pedestrians. The dataset comprises approximately 20 million images from six vehicle-mounted cameras, along with LiDAR, GPS, and INS data, totaling over 20 terabytes.

University dataset [44]: The University dataset, similar to the 7-Scenes dataset, is a challenging benchmark designed for indoor visual localization. It comprises image sequences from five distinct scenes: Office, Meeting, Kitchen, Conference, and Coffee Room. All scenes are registered to a common global coordinate frame, which operates within independent coordinate systems, addressing a key limitation of the 7-Scenes dataset. The dataset includes 9694 training images and 5068 test images.

ApolloScape dataset [152]: The ApolloScape dataset is a large-scale street-view benchmark for tasks such as semantic segmentation, localization, and transfer learning. It consists of over 140,000 RGB video frames with a resolution of 3384×2710 from 73 distinct street-scene videos across China, each annotated with pixel-level semantic masks. Additionally, it includes dense 3D point clouds, as well as depth maps and 3D lane markings categorized into 28 distinct classes. Captured under varying traffic, weather, and lighting conditions, the dataset is divided into easy, moderate, and hard subsets based on scene complexity, such as the density of vehicles and pedestrians.

Aachen Day–Night dataset [153]: The Aachen Day–Night dataset is based on the original Aachen localization dataset; it depicts the old inner city of Aachen, Germany, with a focus on the challenge of localizing nighttime images. It consists of 4328 reference images and 922 query images, including 98 captured at night. The SfM 3D model has 1.65M points. This dataset highlights the difficulty of day-to-night localization, with changes in lighting significantly impacting algorithm performance, making it ideal for testing robustness in urban environments.

RobotCar Seasons dataset [153]: The RobotCar Seasons dataset is based on a subset of the Oxford RobotCar dataset and focuses on localization in urban settings under largely varied seasonal and weather conditions. It includes 20,862 reference images and 11,934 query images captured during different times, such as overcast, sunny, snowy, and rainy weather, as well as at dawn, dusk, and night. The dataset includes a 3D map constructed using SfM with 6.77M 3D points.

CMU Seasons dataset [153]: The CMU Seasons dataset is based on a subset of the CMU dataset and emphasizes seasonal changes and suburban and park scene settings. It contains 7159 reference images and 75,335 query images across 11 traversals of a suburban and park route in Pittsburgh, USA. Seasonal variations include foliage and snow, creating significant challenges in matching features due to dynamic scene geometry changes. The resulting SfM reference model consists of 1.61M 3D points. The dataset provides a comprehensive testbed for evaluating robustness to environmental variability, particularly in outdoor suburban and natural settings.

Table 2. Mainstream 6-DoF pose estimation datasets.

Dataset Name	Scale	Scene	Capture	Challenging Factors	Image Number	3D Map	Point Number	Other Sensor
Vienna [93]	Medium	Urban streets	Car trajectory	Illumination, weather, season	1324 references, 266 queries	SfM	1.12 M	None
Rome [73]	Large	Historic city	Free viewpoint	Occlusion, illumination, people	15,179 references, 1000 queries	SfM	4.31 M	None
CMU [150]	Medium	Urban, suburban, park	Car trajectory	Season, occlusion, illumination	10,000–14,000 per traversal	SfM	-	GPS, IMU
Aachen [94]	Medium	Historic city	Free viewpoint	Day–night, weather, season	3047 references, 369 queries	SfM	1.54 M	None
7-Scenes [131]	Small	Indoor	Free viewpoint	Illumination, motion blur, flat surfaces	26,000 references, 17,000 queries	RGB-D camera	-	None
CLD [108]	Medium	Historic city	Free viewpoint	Illumination, weather, traffic, pedestrian	6848 references, 4081 queries	SfM	1.89 M	None
RobotCar [151]	Large	Urban streets	Car trajectory	Day–night, weather, season, traffic, pedestrian	20M	LiDAR	-	GPS, INS
University [44]	Medium	Indoor	Free viewpoint	Illumination, occlusion	9694 references, 5068 queries	SfM	-	None
ApolloScape [152]	Large	Urban streets	Car trajectory	Illumination, weather, traffic, pedestrian	14K	LiDAR	-	GPS, IMU
Aachen Day–Night [153]	Medium	Historic city	Free viewpoint	Day–night, weather, season	4328 references, 922 queries	SfM	1.65 M	None
RobotCar Seasons [153]	Large	Urban streets	Car trajectory	Day–night, weather, season, traffic, pedestrian	20,862 references, 11,934 queries	SfM	6.77 M	None
CMU Seasons [153]	Medium	Urban, suburban,	Car trajectory	Weather, illumination, season, traffic, pedestrian	7159 references, 75,335 queries	SfM	1.61 M	None

5.2. Evaluation Metrics

To quantitatively evaluate the performance of a monocular camera localization model, three primary indicators are considered: execution time, complexity, and accuracy. However, execution time is notably influenced by the hardware systems employed during

experimentation. This dependency introduces variability, which can skew comparative analyses. In addition, a limited number of studies disclose the time and space complexity data of their proposed methods. To sidestep the inconsistencies introduced by varying studies, this paper prioritizes accuracy evaluation metrics to ensure a fair and valid benchmarking of different methodologies.

Place recognition and 6-DoF pose estimation, as two key components of monocular camera localization, require distinct accuracy metrics: the former relies on image retrieval performance, while the latter is assessed based on position and orientation errors. These key metrics are detailed as follows.

5.2.1. Place Recognition Metrics

Mean Average Precision (mAP): The standard evaluation metric used in image retrieval and place recognition tasks is mAP. For each query, the system retrieves a ranked list of images from a dataset, which can be used to calculate an Average Precision (AP)—the area under the precision–recall curve. The precision (P) and recall (R) at each rank k in the retrieved list are defined as

$$P(k) = \frac{\text{Number of correctly retrieved images up to rank } k}{k} \quad (2)$$

$$R(k) = \frac{\text{Number of correctly retrieved images up to rank } k}{\text{Total number of relevant images in the dataset}} \quad (3)$$

The Average Precision (AP) for a single query is the weighted sum of precision values at ranks where a relevant image is retrieved, considering only the changes in recall:

$$AP = \sum_{k=1}^N P(k) \cdot [R(k) - R(k-1)] \quad (4)$$

where $R(k) - R(k-1)$ represents the change in recall between consecutive ranks.

An ideal precision–recall curve has precision 1 over all recall levels, and this corresponds to an AP of 1. By averaging the AP scores of multiple queries, a single mAP number can be used to evaluate the overall performance.

Finally, the mAP is obtained by averaging the AP values across all queries:

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP_q \quad (5)$$

where Q is the total number of queries, and AP_q is the Average Precision for the q -th query.

Mean Average precision at rank K (mAP@K): In addition to the mAP metric, some studies also report the mAP at rank K. The former evaluates the entire retrieved list, while the latter only considers the top K ranked images in all retrieved lists. Specifically, for mAP, the value of N in Equation (2) equals the length of the retrieved list, while for mAP@K, N equals a fixed number, K . This metric is sometimes preferred over mAP since exhaustive retrieval of every matching image is computationally expensive and may not be necessary in some applications.

5.2.2. 6-DoF Pose Estimation Metrics

The two most used metrics in 6-DoF pose estimation are the translation and rotation errors.

Translation Error (TE): The TE measures the Euclidean distance between the estimated position vector (p_e) and the ground truth position vector (p_{gt}) in a local x, y, and z (or east, north, and up) axis:

$$TE = \|p_e - p_{gt}\| \quad (6)$$

TE quantifies how accurately the system estimates the spatial position of the camera, typically measured in meters or centimeters.

Rotation Error (RE): The RE quantifies the minimum angular difference between the estimated orientation and the ground truth orientation:

$$RE = 2 \cdot \cos^{-1}(|q_e - q_{gt}|) \quad (7)$$

where q_e and q_{gt} are the estimated and ground truth orientations in quaternions.

RE evaluates how well the orientation is estimated. It is calculated in radians and is typically translated to degrees ($^{\circ}$).

Average Recall (AR): In addition to TE and RE, some studies also report AR, which measures the percentage of poses that meet specific error thresholds for both TE and RE:

$$AR = \frac{\text{Number of correct poses}}{\text{Total number of poses}} \quad (8)$$

where an estimated pose is correct only if both TE and RE are within predefined thresholds.

6. Comparative Analysis

This section evaluates the accuracy performance of camera localization methods on several of the most widely used place recognition and 6-DoF pose estimation datasets. We focus on methods demonstrating competitive performance at the time of publication. Due to varying evaluation practices and limited source code availability, some datasets may include a limited selection of models. By examining the performance metrics of these models in relation to different dataset characteristics and model architectures, we aim to provide a comprehensive overview of the capabilities and limitations of different approaches.

6.1. Comparing Place Recognition Methods on Common Benchmarks

We evaluate the mAP accuracy of the place recognition techniques on the revisited Oxford and Paris datasets ($\mathcal{ROxford}$ and $\mathcal{RParis}$), as shown in Tables 3 and 4, as well as the original Oxford and Paris datasets, as shown in Table 5. $\mathcal{ROxford}$ and $\mathcal{RParis}$ have become the most commonly used benchmarks for evaluating recent place recognition methods. Each of them features medium and hard difficulty levels. Medium includes queries with moderate viewpoint and appearance changes, while hard features significant viewpoint shifts, extreme lighting variations, and heavy occlusions. Both levels are evaluated with the original set (relevant images only), plus the R1M set, which adds one million non-relevant images, largely increasing evaluation complexity. Conversely, the original Oxford and Paris datasets are primarily used to evaluate earlier methods. They feature original versions (Oxford5K and Paris6K) and extended versions (Oxford105K and Paris106K), which introduce 100,000 distractor images to the original datasets.

Table 3. mAP (%) scores of place recognition methods evaluated on the $\mathcal{ROxford}$ dataset. Results are shown for medium and hard difficulty levels, with and without the R1M distractor set.

Method	Medium		Hard	
	$\mathcal{ROxford}$	$\mathcal{ROxford} + \text{R1M}$	$\mathcal{ROxford}$	$\mathcal{ROxford} + \text{R1M}$
SuperGlobal [38]	90.90	84.40	80.20	71.10
Hypergraph Propagation + Community Selection [31]	88.40	79.10	73.00	60.50
Tokenizer [36]	82.28	75.64	66.57	51.37
FIRe [37]	81.80	66.50	61.20	40.10
DOLG [35]	81.50	77.43	58.82	52.21
RRT + αQE [39]	80.40	71.70	64.00	50.90
HOW [32]	79.40	65.80	56.90	38.90
DELG [33]	78.50	62.70	59.30	39.30
NetVLAD [24]	73.91	60.51	56.45	37.92
GEM + αQE [29]	71.40	53.10	45.90	26.20
DELf-ASMK + SP [27]	67.80	53.80	43.10	31.20
GEM [29]	64.70	45.20	38.50	19.90
R-MAC [21]	60.90	39.30	32.40	12.50

Table 4. mAP (%) scores of place recognition methods evaluated on the $\mathcal{RParis}$ dataset. Results are shown for medium and hard difficulty levels, with and without the R1M distractor set.

Method	Medium		Hard	
	$\mathcal{RParis}$	$\mathcal{RParis} + \text{R1M}$	$\mathcal{RParis}$	$\mathcal{RParis} + \text{R1M}$
SuperGlobal [38]	93.30	84.90	86.70	71.40
Hypergraph Propagation + Community Selection [31]	92.60	86.60	83.30	72.70
DOLG [35]	89.81	80.79	77.70	62.83
Tokenizer [36]	89.34	79.76	78.56	61.56
RRT + αQE [39]	88.50	74.80	77.70	57.10
NetVLAD [24]	86.81	71.31	73.61	48.98
FIRe [37]	85.30	67.60	70.00	42.90
GEM + αQE [29]	84.00	60.30	67.30	32.30
DELG [33]	82.90	62.60	65.50	37.00
HOW [32]	81.60	61.80	62.40	33.70
R-MAC [21]	78.90	54.80	59.40	28.00
GEM [29]	77.20	52.30	56.30	24.70
DELf-ASMK + SP [27]	76.90	57.30	55.40	26.40

Table 5. mAP (%) scores of place recognition methods evaluated on the Oxford5K, Oxford105K, Paris6K, and Paris106K datasets.

Method	Oxford5K	Oxford105K	Paris6K	Paris106K
E2E-R-MAC + QE [28]	90.60	89.40	96.00	93.20
DIR +QE [23]	87.10	85.20	95.30	91.80
DIR [23]	86.10	82.80	94.50	90.60
MAC + QE [25]	85.00	81.80	86.50	78.80
DELf [27]	83.80	82.60	85.00	81.70
R-MAC + QE [21]	82.90	77.90	85.60	78.30
BoW (iSP + ctx QE) [16]	82.70	76.70	80.50	71.00
BoW (SPAUG + DQE + RootSIFT) [17]	80.90	72.20	-	-
MAC [25]	79.70	73.90	82.40	74.60

BoW (Elliptical regions + SP + QE) [14]	78.40	72.80	-	-
R-MAC [21]	77.00	69.20	83.80	76.40
CroW + QE [26]	74.90	70.60	84.80	79.40
CroW [26]	70.80	65.30	79.70	72.20
BoW (tf-idf + SP) [11]	67.20	58.10	-	-
SPoC [22]	53.10	-	50.10	-

ROxford and RParis. In Tables 3 and 4, SuperGlobal, Hypergraph Propagation with Community Selection, DOLG, and Tokenizer consistently achieve the highest mAP scores across both the ROxford and RParis datasets. SuperGlobal ranks first on both datasets with its highly optimized global feature extraction framework, which uses regional and scale-specific pooling (Regional-GeM and Scale-GeM) to effectively balance compactness and discriminative power. Hypergraph Propagation with Community Selection follows closely, leveraging graph-based models to propagate spatial relationships and reduce feature matching ambiguity. Furthermore, DOLG and Tokenizer exhibit competitive performance due to their ability to integrate local and global feature learning seamlessly. DOLG integrates complementary local and global feature representations using orthogonal fusion and applies attention mechanisms to prioritize spatially significant features. Tokenizer, on the other hand, aggregates local features into visual tokens using self-attention and cross-attention mechanisms, enabling efficient and semantically rich representations.

HOW, DELG, and NetVLAD place in the middle tier. HOW innovates by combining learned deep local descriptors with ASMK for efficient and effective matching, providing moderate performance in structured environments. DELG unifies global and local features, employing generalized mean pooling for global descriptors and attentive selection for local features. NetVLAD, one of the foundational methods in this domain, integrates VLAD pooling into a CNN framework for end-to-end training.

In contrast, GeM, R-MAC, and DELF-ASMK + SP rank lower. GeM, with its generalized mean pooling approach, provides a simple but less discriminative global descriptor. R-MAC aggregates regional features but may struggle to capture fine-grained spatial details. DELF-ASMK + SP, though benefiting from spatial verification, remains constrained by its reliance on handcrafted aggregation techniques.

Oxford and Paris. E2E-R-MAC, DIR, and DELF rank as the top-performing methods. E2E-R-MAC stands out with its end-to-end training pipeline, which integrates regional pooling and dimensionality reduction into a unified framework. DIR builds on the R-MAC framework by incorporating region proposal networks and triplet loss, providing content-aware pooling and improved regional feature quality, which results in high precision and robustness. DELF, although ranking lower compared to more recent methods, demonstrated strong performance in its early years.

Medium-performing methods include R-MAC, MAC, BoW (iSP + ctx QE), BoW (SPAUG + DQE + RootSIFT), and BoW (Elliptical regions + SP + QE). While R-MAC ranks lower compared to more recent methods on the ROxford and RParis datasets, it demonstrated moderate performance during its early years. MAC also achieves competitive results by employing maximum activation pooling to create compact image representations. The BoW-based methods in this category improve traditional BoW approaches through refinements such as incremental Spatial Re-Ranking (iSP), Contextual Query Expansion (ctx QE), spatial augmentation (SPAUG), Discriminative Query Expansion (DQE), and RootSIFT features, making them competitive compared to learning-based approaches.

Lower-performing methods include CroW, SPoC, and BoW (tf-idf + SP). Despite the early adoption of deep convolutional features from SPoC, its use of a simplistic sum pooling strategy may not be able to capture complex image details effectively. BoW (tf-idf + SP) incorporates frequency-inverse document frequency (tf-idf) weighting and spatial

pooling. Yet, its heavy reliance on handcrafted features and visual word representations may still struggle with complex scenes.

QE and α QE improvements. QE (query expansion) refines image retrieval results by iteratively updating the query representation based on the top-ranked retrieved images. This dynamic adjustment enhances recall and precision. α QE extends QE by assigning weighted contributions to the top-ranked images based on their similarity to the query. This α -weighting mechanism ensures that highly similar results have a stronger influence on the refined query.

In Tables 3 and 4, GEM + α QE shows improved accuracy over GEM due to the α -weighted refinement process. In Table 5, the incorporation of QE enhances multiple methods: DIR + QE outperforms DIR, MAC + QE outperforms MAC, R-MAC + QE outperforms R-MAC, and CroW + QE outperforms CroW. These enhancements demonstrate the consistent benefits of QE and α QE across diverse retrieval methods.

6.2. Comparing 6-Dof Pose Estimation Methods on Common Benchmarks

This section evaluates the six major classes of 6-DoF pose estimation methods across widely recognized benchmarks—7-Scenes, CLD, Aachen Day–Night, and RobotCar Seasons datasets—as shown in Tables 6–9, respectively. The 7-Scenes and CLD datasets are the most established benchmarks for APR and SCR techniques, providing diverse indoor and outdoor scenarios, respectively. They primarily evaluate performance using TE and RE metrics. Conversely, Aachen Day–Night and RobotCar Seasons serve as mainstream benchmarks for feature matching-based localization, offering urban- and street-level scenes under varying conditions, such as lighting, weather, and occlusions. These datasets assess localization robustness using AR at three precision thresholds: high ($\leq 0.25\text{m}$, $\leq 2^\circ$), medium ($\leq 0.5\text{m}$, $\leq 5^\circ$), and low ($\leq 5\text{m}$, $\leq 10^\circ$). The following presents intra-class comparisons, which analyze the performance of methods within each localization class, and inter-class comparisons, which examine performance trends across the six major localization classes.

Table 6. TE and RE of localization methods evaluated on the 7-Scenes indoor dataset, averaged across the seven scenes: chess, fire, heads, office, pumpkin, kitchen and stairs. Note: Lower values indicate better performance.

Method Type	Method	TE (cm)	RE (°)
Approximate 2D-2D	DenseVLAD [20]	26.14	13.11
Precise 2D-2D	CamNet [48]	3.86	1.69
	AnchorNet [47]	9.00	6.74
	EssNet [50]	19.00	4.28
	RelocNet [46]	21.00	6.73
	Relative PN [44]	21.00	9.28
Direct 2D-3D	PixLoc [88]	2.86	0.98
	Active Search [79]	3.71	1.18
Hierarchical 2D-3D	HF-Net [100]	3.14	1.09
	InLoc [98]	4.14	1.38
APR	VLocNet++ [125]	2.15	1.39
	DFNet+ [130]	2.42	0.79
	VLocNet [124]	4.80	3.80
	DFNet [129]	6.42	1.93
	MS-Transformer+ [122]	15.00	7.28
	MS-Transformer [121]	18.00	7.28
	MLF [128]	18.42	7.44
	AdPR [126]	19.42	7.47

SCR	AtLoc [116]	19.71	7.56
	MapNet [115]	20.71	7.79
	GeoPoseNet [110]	22.86	8.12
	HourGlass-Pose [112]	23.29	9.53
	VidLoc [123]	25.00	-
	BranchNet [113]	29.00	8.30
	LSTM PoseNet [111]	31.29	9.85
	PoseNet [108]	44.14	10.40
	Bayesian PoseNet [109]	46.57	9.81
	ACE++ [145]	0.30	1.00
	ACE [144]	0.33	1.08
	HSCNet++ [139]	2.29	0.81
	HSCNet [138]	2.71	0.90
	DSAC* [137]	2.71	1.36
	NeuMap [141]	3.14	1.09
	MaRepo [142]	3.18	1.54
	DSAC++ [135]	3.57	1.10
	SACReg [143]	3.71	1.22
	DSAC [134]	20.00	6.30

Table 7. TE and RE of localization methods evaluated on the CLD outdoor dataset, averaged across four landmark categories: King’s College, Old Hospital, St. Mary’s Church, Shop Façade. Note: Lower values indicate better performance.

Method Type	Method	TE (cm)	RE (°)
Approximate 2D-2D	DenseVLAD [20]	255.75	7.10
Precise 2D-2D	EssNet [50]	83.00	1.36
	AnchorNet [47]	84.00	2.10
Direct 2D-3D	BGNet [86]	5.93	0.13
	FuseLoc [92]	10.00	0.20
	UM [83]	11.23	0.40
	Active Search [79]	13.80	0.23
	PixLoc [88]	15.00	0.25
Hierarchical 2D-3D	HF-Net [100]	10.80	0.20
APR	DFNet+ [130]	35.25	0.77
	VLocNet [124]	78.40	2.82
	MS-Transformer+ [122]	96.00	2.73
	DFNet [129]	119.25	2.90
	MS-Transformer [121]	128.00	2.73
	LSTM PoseNet [111]	130.00	5.52
	SVS-Pose [114]	132.50	5.17
	GeoPoseNet [110]	163.25	2.86
	Bayesian PoseNet [109]	192.00	6.28
	PoseNet [108]	208.50	6.83
	SACReg [143]	8.75	0.23
SCR	ACE [144]	10.25	0.30
	ACE++ [145]	11.25	0.28
	HSCNet [138]	13.00	0.30
	HSCNet++ [139]	13.50	0.29
	DSAC* [137]	13.50	0.35
	NeuMap [141]	14.00	0.33
	DSAC++ [135]	14.25	0.33
	DSAC [134]	31.75	0.78

Table 8. AR of localization methods evaluated on the Aachen Day–Night dataset. Three different AR thresholds are used: high-precision (0.25 m, 2°), mid-precision (0.5 m, 5°), and low-precision (5 m, 10°). Note: Higher values indicate better performance.

Method Type	Method	Day			Night		
		High	Mid	Low	High	Mid	Low
Approximate 2D-2D	FAB-MAP [13]	0.0	0.0	4.6	0.0	0.0	0.0
	NetVLAD [24]	0.0	0.2	18.9	0.0	0.0	14.3
	DenseVLAD [20]	0.0	0.1	22.8	0.0	1.0	19.4
Direct 2D-3D	Active Search [79]	85.3	92.2	97.9	39.8	49.0	64.3
	BGNet [86]	84.5	92.4	96.2	46.9	63.3	84.7
	PixLoc [88]	84.7	94.2	98.8	81.6	93.9	100.
Hierarchical 2D-3D	UR2KiD [103]	79.9	88.6	93.6	45.9	64.3	83.7
	S2DNet [102]	84.5	90.3	95.3	74.5	82.7	94.9
	D2-Net [101]	84.3	91.9	96.2	75.5	87.8	95.9
	DSLod [104]	89.3	95.4	97.6	44.9	67.3	87.8
	HF-Net [100]	89.6	95.4	98.8	86.7	93.9	100.
SCR	DSAC [134]	0.4	2.4	34.0	-	-	-
	ESAC [136]	42.6	59.6	75.5	6.1	10.2	18.4
	HSCNet [138]	65.5	77.3	88.8	22.4	38.8	54.1
	NeuMap [141]	76.2	88.5	95.5	37.8	62.2	87.8
	SACReg [143]	85.8	95.0	99.6	67.5	90.6	100.

Table 9. AR of localization methods evaluated on the RobotCar Seasons dataset. Three different AR thresholds are used: high-precision (0.25 m, 2°), med-precision (0.5 m, 5°), and low-precision (5 m, 10°). Note: Higher values indicate better performance.

Method Type	Method	Day			Night		
		High	Mid	Low	High	Mid	Low
Approximate 2D-2D	NetVLAD [24]	6.4	26.3	90.9	0.3	2.3	15.9
	DenseVLAD [20]	7.6	31.2	91.2	1.0	4.4	22.7
Direct 2D-3D	Active Search [79]	50.9	80.2	96.6	6.9	15.6	31.7
	BGNet [86]	55.7	77.0	94.1	8.4	21.7	50.9
	PixLoc [88]	56.9	82.0	98.1	34.9	67.7	89.5
Hierarchical 2D-3D	S2DNet [102]	53.9	80.6	95.8	14.5	40.2	69.7
	D2-Net [101]	54.5	80.0	95.3	20.4	40.1	55.0
	HF-Net [100]	56.9	81.7	98.1	33.3	65.9	88.8

6.2.1. Intra-Class Comparisons

Comparing APR methods. In Tables 6 and 7, DFNet+ consistently achieves the top TE and RE on both datasets. This demonstrates its superior capability to handle diverse indoor and outdoor conditions, which can be attributed to its neural feature synthesis, which refines pose predictions through adaptive color transformations and effective feature fusion. Next, VLocNet++ and VidLoc are close competitors. The strong performance of these methods is driven by pose refinements through spatio-temporal consistency, which leverages sequences of frames rather than single images. Subsequently, MS-Transformer and MS-Transformer+ achieve high rankings, benefiting from transformer-based architectures that enhance the extraction of position and orientation features, along with scene-specific attention mechanisms.

Mid-ranking methods such as AtLoc, AdPR, and MapNet show moderate improvement over early PoseNet variants. AtLoc utilizes attention mechanisms to prioritize stable geometric regions in images, which likely accounts for its robustness in variable scenes. AdPR provides competitive results by leveraging adversarial training to model the joint distribution of RGB images and camera poses. MapNet also performs relatively well, aided by incorporating geometric constraints and sequential optimization to enhance pose estimation.

In contrast, early methods such as PoseNet, Bayesian PoseNet, LSTM PoseNet, and GeoPoseNet exhibit the highest errors. PoseNet's basic CNN-based architecture limits its ability to handle complex scenes. Bayesian PoseNet incorporates uncertainty modeling, which enhances prediction accuracy. LSTM PoseNet incorporates sequential data processing through recurrent structures, which offers better handling of dynamic environments. GeoPoseNet enhances PoseNet's framework by incorporating geometric constraints into the loss function.

Comparing SCR methods. ACE and ACE++ emerge as the top-performing methods on both datasets, benefiting from their efficient architecture combining Accelerated Coordinate Encoding and attention mechanisms, which dynamically refine important features. Following closely, the HSCNet series demonstrates strong performance by utilizing hierarchical coarse-to-fine prediction frameworks, with HSCNet++ benefiting from transformer-based modules in capturing global scene contexts, combined with hierarchical pseudo-labeling and an angle-based reprojection loss.

SACReg, on the other hand, exhibits significantly different performance values across the two benchmarks. SACReg excels in the CLD outdoor dataset, where its scene-agnostic design, powered by vision transformers and cosine-based coordinate encoding, effectively generalizes to complex scenes with significant variations. Conversely, on the 7-Scenes indoor dataset with simpler or more structured environments, SACReg's scene-

agnostic design may not leverage dataset-specific nuances as effectively as methods like ACE++ or HSCNet++, resulting in relatively low performance.

The DSAC series, including DSAC, DSAC++, and DSAC*, ranks as the worst-performing models. While DSAC introduces the differentiable sample consensus mechanism, its initial implementation falls behind compared to more recent methods with global-context-awareness and advanced feature extraction capabilities. DSAC++ improves this approach by incorporating a soft inlier count mechanism, which enhances generalization and reduces overfitting. DSAC* further improved accuracy by simplifying the training procedure and updating the backbone network architecture.

Comparing approximate 2D-2D localization methods. Unlike other localization techniques, approximate 2D-2D methods are primarily evaluated on image retrieval datasets, as highlighted in Section 6.1. This focus on place recognition datasets limits the availability of evaluation results on 6-DoF pose estimation benchmarks, such as those featured in Tables 6–9. Despite this limitation, we compare the performance of FAB-MAP, NetVLAD, and DenseVLAD in Tables 8 and 9.

DenseVLAD achieves superior accuracy among the three methods. Its dense sampling strategy for local descriptors, combined with VLAD aggregation, ensures strong representation of both local and global image features. NetVLAD follows DenseVLAD closely in performance by integrating VLAD pooling into a CNN architecture for end-to-end feature learning. FAB-MAP ranks lowest, relying on a probabilistic approach to model co-occurrence relationships between visual words. While effective in mitigating perceptual aliasing and dynamic environments, its reliance on handcrafted feature descriptors may limit its capacity to handle the complexity and variability.

Comparing precise 2D-2D localization methods. Precise 2D-2D localization methods tackle 6-DoF pose estimation by refining the initial result from approximate 2D-2D localization using relative camera pose estimation. In Table 6, CamNet emerges as the top-performing method across the benchmarks, reflecting its hierarchical triple-step approach, which separates the tasks of image-based coarse retrieval, pose-based fine retrieval, and precise relative pose regression into distinct modules. AnchorNet follows as a strong performer with its novel anchor-based approach inspired by human navigation—assigning anchor points across the scene and predicting relative offsets. EssNet ranks third by utilizing an essential matrix-based approach to directly estimate relative poses from image pairs. Earlier approaches like RelocNet and Relative PN, while foundational, rely on simpler architectures and lack advanced features, such as hierarchical processing or geometric loss functions seen in higher-ranking methods, resulting in relatively low accuracy.

Comparing direct 2D-3D localization methods. In Tables 6, 8, and 9, PixLoc emerges as a top performer on the 7-Scenes, Aachen Day–Night, and RobotCar Seasons datasets. Notably, it performs exceptionally well on the nighttime subsets of the Aachen Day–Night and RobotCar Seasons datasets. This performance is likely benefited by its scene-agnostic design, enabled by a multi-scale feature extraction pipeline and differentiable pose refinement, enabling scene-agnostic predictions. The incorporation of uncertainty-based weighting to focus on stable scene elements makes it robust in dynamic and ambiguous environments. However, PixLoc performs poorly on the CLD dataset, as shown in Table 7. This discrepancy likely stems from the unique characteristics of the CLD dataset, which involves large-scale urban environments with highly structured and repetitive elements. While PixLoc’s scene-agnostic design excels in generalization, it lacks specialized mechanisms to handle dense, repetitive structures effectively.

In the next rank, BGNet demonstrates strong performance by employing a bidirectional geometric network, which leverages both geometric constraints and context-aware attention mechanisms for precise feature matching. Active Search ranks as the worst-

performing method overall, despite its efficiency in large-scale structured datasets. Its reliance on handcrafted features and traditional optimization techniques likely limits its accuracy and adaptability in diverse or complex environments compared to learning-based methods.

Comparing hierarchical 2D-3D localization methods. Hierarchical 2D-3D localization methods reduce the computational costs of direct 2D-3D methods by narrowing down the portion of the 3D model involved in matching, achieving a balance between efficiency and accuracy. Tables 8 and 9 compare the performance of hierarchical methods like HF-Net, DSLoc, D2-Net, S2DNet, and UR2KiD.

Overall, HF-Net ranks as the top performer on both datasets with its hybrid approach of combining global descriptors for image retrieval with local descriptors for precise matching. Next, DSLoc demonstrates similar performance on the daytime subset of the Aachen Day–Night dataset, where its hybrid approach combining handcrafted and learned features with dense semantic 3D maps excels. However, its performance on the nighttime subset drops considerably. D2-Net ranks next by leveraging dense feature extraction and detect-and-describe optimization for robust 2D-3D matching. S2DNet follows closely, employing a sparse-to-dense matching approach that frames feature matching as a supervised classification task. UR2KiD ranks last among the hierarchical methods, with particularly poor performance on the nighttime subset of the Aachen Day–Night dataset. This may be caused by its reliance on geometric priors and scene-specific filtering, which fail to generalize to nighttime scenes.

6.2.2. Inter-Class Comparisons

The evaluation results of the six types of localization methods—approximate and precise 2D-2D localization, direct and hierarchical 2D-3D localization, and regression-based methods (SCR and APR) are compared as follows.

2D-2D Localization Methods. Approximate 2D-2D methods generally show the lowest accuracy across datasets. This limitation arises from their reliance on the retrieved image pose as the final estimated pose, without geometric refinement. Precise 2D-2D methods, by incorporating a relative pose estimation step, improve upon approximate methods. However, even with this enhancement, precise 2D-2D methods exhibit the lowest overall performance when compared to 2D-3D localization and regression-based methods.

2D-3D Localization Methods. Direct and hierarchical 2D-3D localization methods demonstrate comparable accuracy across datasets. Direct methods establish 2D-3D correspondences directly, offering high accuracy but often at the cost of computational efficiency. Hierarchical 2D-3D methods optimize the search space in the 3D model using retrieval-based pre-selection, achieving more efficient computation. Both types of 2D-3D methods generally outperform regression-based methods.

Regression-Based Methods. Among regression-based approaches, SCR methods consistently outperform APR methods. SCR methods predict dense 3D scene coordinates for each image pixel, enabling precise pose estimation through geometric refinement. This allows them to surpass the accuracy of APR methods, which rely on direct pose regression. While recent SCR methods have achieved performance comparable to 2D-3D methods on outdoor benchmarks, SCR methods such as ACE and ACE++ exhibit superior performance on the 7-Scenes indoor dataset.

General Trends. Generally, 2D-3D localization methods tend to achieve the highest accuracy among all classes. However, recent advancements in SCR methods, such as ACE and ACE++, have reached comparable and, in indoor scenes, even superior performance to 2D-3D methods, all while maintaining simpler pipeline designs. This demonstrates the potential of SCR methods to balance high accuracy with computational efficiency, making

them increasingly competitive. In comparison, precise 2D-2D methods, while offering improved accuracy over approximate 2D-2D methods, achieve the lowest accuracy among the major localization approaches. This may show the importance of leveraging 3D information or advanced regression-based techniques to enhance localization precision.

7. Challenges and Future Work

Monocular camera localization in known environments has seen significant advancements, driven by the development of robust feature extraction techniques, efficient 2D-2D and 2D-3D matching methods, and deep learning-based pose estimation frameworks. Despite these achievements, the introduction of new technologies invariably brings novel challenges. In this section, we identify several critical issues and propose future research directions.

7.1. Synthetic Data Generation

State-of-the-art deep learning-based monocular localization methods often require large volumes of high-quality training data to achieve impressive performance. However, creating such datasets is a significant challenge due to the time, cost, and effort required for large-scale annotation and data collection. Despite the creation of datasets like Aachen Day–Night and RobotCar Seasons, which cover diverse and large-scale scenes with varied lighting and seasonal conditions, the fast-paced advancements in highly capable deep learning models continuously demand even larger datasets with higher variability. This creates a persistent gap between dataset availability and the needs of cutting-edge models, limiting their ability to push the boundaries of monocular camera localization.

To address this challenge, recent research has turned to synthetic data generation as a promising solution. Techniques like GAN-based image synthesis enable the automatic generation of training data, including edge cases such as extreme lighting conditions or rare environmental configurations, which are difficult to capture in real-world datasets. For example, the Dark Side Augmentation approach [41] employs a CycleGAN-based architecture [154] to synthesize realistic nighttime images from daytime data, significantly improving retrieval performance in low-light scenarios without requiring real nighttime annotations. Similarly, the recent DFNet+ model [11] demonstrates the value of integrating synthetic views—generated via neural rendering techniques—to enhance model robustness against viewpoint changes and occlusions. To this end, we recognize the potential of synthetic data generation to enhance the robustness of monocular localization methods.

Future research pathways could focus on several open sub-problems to advance this area:

- Developing specialized GAN variants, such as geometry-aware GANs, that preserve 3D structural consistency during synthesis to better support tasks like scene coordinate regression (e.g., ensuring synthesized images maintain accurate depth cues for 2D-3D matching);
- Exploring hybrid synthesis methods that combine physics-based rendering (e.g., via tools like Blender or Unreal Engine) with GANs to simulate dynamic elements like weather or crowd movements, which are underrepresented in current datasets;
- Addressing scalability issues in generating diverse, large-scale synthetic environments by automating procedural generation pipelines tailored to urban or indoor scenes.

Additionally, research efforts should focus on domain adaptation methods to reduce the domain gap between synthetic and real-world data, ensuring that models trained on augmented or synthetic datasets perform reliably in practical applications. Concrete

avenues include adversarial domain adaptation frameworks [155] to align feature distributions across domains or self-supervised techniques like contrastive learning [156] to fine-tune models on unlabeled real data.

7.2. Generalizable Models

One of the key limitations for mainstream 2D-3D and SCR localization methods is their limited generalization ability across diverse environments. These deep learning-based methods often excel on specific datasets but struggle when applied to unseen scenes with significant changes in appearance, scale, or environmental conditions. Extensive fine-tuning is frequently required to adapt these models to new settings.

To address these limitations, future research should prioritize the development of models that generalize better across diverse environments without the need for extensive retraining. Recent methods have begun to adopt this trend. For example, PixLoc [88] demonstrates scene-agnostic performance by optimizing multi-scale feature alignment with differentiable pose refinement, reducing reliance on scene-specific training. Similarly, SACReg [138] leverages a Vision Transformer architecture combined with cosine-based coordinate encoding and cross-attention mechanisms to generalize to diverse environments without embedding scene-specific information into model weights. Additionally, hybrid frameworks like MaRepo [137], which combine learned feature representations with geometric reasoning, provide robustness to variations in environmental conditions. Building on these advancements, future research pathways could focus on several open sub-problems to further enhance generalization:

- Exploiting permutation equivariance in relative pose estimation, as in EquiPose [157], to handle unordered correspondences and improve zero-shot performance across varying scene structures;
- Incorporating hypernetworks for adaptive localization, such as in HyperPose [158], which uses meta-learning to dynamically adjust parameters for unseen environments and extends benchmarks like Cambridge Landmarks for broader evaluation;
- Scaling relative pose regression through large-scale training, as explored in Reloc3r [159], integrating foundation models pre-trained on massive datasets to enable fast, accurate localization with minimal adaptation, addressing domain shifts in real-time applications like autonomous navigation.

7.3. Multi-Sensor Fusion

Monocular camera localization methods primarily rely on visual information, which can face challenges in applications requiring fail-safe localization, such as autonomous driving, train positioning, and drone navigation. A promising direction for addressing these limitations is the integration of multi-sensor fusion, combining visual data with other modalities such as GPS, IMU, LiDAR, and RADAR. For applications requiring high reliability, these additional sensors provide complementary information that enhances the robustness and accuracy of localization systems. For instance, GPS can offer coarse pose constraints, while IMU data can improve orientation estimation during rapid movements or occlusions. LiDAR and RADAR add geometric and depth information, helping to resolve ambiguities in complex or texture-less environments.

Research could further explore fusion strategies at various levels:

1. **Data-Level Fusion:** Combining raw sensor data, such as aligning LiDAR point clouds with camera images to create dense spatial representations, as seen in BEVFusion [160], which unifies Bird's-Eye-View representations for multi-task perception.

2. Feature-Level Fusion: Integrating feature descriptors from different modalities to improve matching accuracy, for instance, using cross-modal attention in transformer architectures to align tokens from camera and LiDAR embeddings.
3. Decision-Level Fusion: Merging individual pose estimates from various sensors for improved robustness and consistency, often enhanced by probabilistic models or Kalman filters to handle uncertainties.

In addition, future research pathways could focus on several open sub-problems to advance multi-sensor fusion:

- Developing adaptive fusion mechanisms that dynamically weigh sensor contributions based on environmental conditions, extending methods like SAMFusion [161] to incorporate real-time uncertainty estimation via Bayesian transformers;
- Addressing spatio-temporal misalignment in dynamic scenes by exploring Mamba-style recurrent architectures [162–164] for efficient sequential fusion, potentially improving upon current transformer models in computational efficiency for edge devices;
- Leveraging large vision–language models for semantic-aware fusion, such as integrating CLIP-like embeddings to enhance cross-modal alignment in unstructured environments, thereby improving generalization across diverse datasets like nuScenes or Waymo Open.

7.4. Advanced Learning Paradigms

While significant progress has been made in monocular camera localization, most methods rely on conventional deep learning architectures such as CNNs for feature extraction and pose estimation. These architectures, while effective, often struggle with capturing long-range dependencies and contextual relationships across diverse scenes.

Recent top-performing models have shown a significant trend toward incorporating Vision Transformer architectures. For example, SACReg [143], HSCNet++ [139], MS-Transformer+ [122], and NeuMap [141] achieve superior accuracy by leveraging self-attention mechanisms to effectively capture global context and long-range dependencies. These models outperform traditional CNN-based approaches, demonstrating exceptional robustness in handling complex and diverse environments. However, transformers tend to be computationally intensive, which necessitates future research into designing lightweight and efficient transformer-based models suitable for real-time applications.

Beyond transformers, emerging techniques like GNNs are showing promise in modeling spatial and geometric relationships. For instance, BGNet [86] employs bipartite graph structures to improve correspondence prediction in 2D-3D matching tasks, enhancing robustness in structured and repetitive environments compared to CNN-based methods. Recent advancements, such as the attentional GNN for massive network localization [165], further demonstrate GNNs' potential in scalable visual relocalization.

In the future, several potential research pathways could be focused on to advance this area:

- Integrating state–space models like Mamba into vision pipelines, as in VimGeo [166], to achieve linear-time complexity for long-sequence processing in dynamic environments while maintaining accuracy;
- Developing hybrid architectures that combine transformers, GNNs, and SSMs for multi-modal fusion, addressing computational bottlenecks in real-time applications;
- Exploring diffusion-based paradigms for handling pose uncertainty, inspired by promptable 3D localization models [167], to improve robustness in noisy or incomplete data scenarios.

7.5. Lightweight Architectures

State-of-the-art monocular camera localization methods largely rely on complex neural network architectures and high-dimensional feature representations to achieve impressive accuracy but at the cost of high computational complexity and memory requirements. This poses a significant challenge for deploying these methods on resource-constrained devices, such as mobile phones, drones, and autonomous robots, and for real-time applications.

Developing lightweight architectures that maintain high accuracy while minimizing computational demands is a crucial research direction. Recent studies have shed light on this challenge. For example, AnchorNet [47] demonstrates how focusing on anchor points can reduce runtime and memory requirements and simplify the localization pipeline without significant loss of accuracy. Similarly, HF-Net [100] leverages shared computation across global and local feature extraction, enhancing efficiency while maintaining robustness in diverse environments. However, the portability and real-time applicability of these highly precise localization methods remain ongoing challenges.

Potential future research pathways to advance this area include:

- Applying advanced model compression techniques, such as dynamic pruning and 4-bit quantization, to further reduce the footprint of transformer-based models while preserving pose estimation accuracy in dynamic scenes;
- Exploring hybrid CNN-Mamba architectures [168–170] to handle long-range dependencies efficiently without the quadratic complexity of self-attention, potentially integrating with datasets like Euroc MAV for UAV-specific evaluations;
- Developing adaptive lightweight models that switch between low-power modes based on device constraints, incorporating federated learning for on-device personalization in diverse environments like urban navigation.

8. Conclusions

Monocular camera localization in known environments plays a crucial role in enabling reliable and efficient positioning for various applications, such as autonomous navigation, augmented reality, and robotics. This review provided a comprehensive analysis of the fundamental techniques and recent advancements in the field, categorized into 2D-2D feature matching-based methods (approximate 2D-2D localization and precise 2D-2D localization), 2D-3D feature matching-based methods (direct 2D-3D localization and hierarchical 2D-3D localization), and regression-based approaches (APR and SCR). Comparative analysis was conducted with both inter-class and intra-class comparisons to evaluate the performance differences within and across these categories on mainstream benchmarks.

The findings indicate that 2D-3D localization methods generally achieve the highest accuracy among all classes. However, recent advancements in SCR methods, such as ACE and ACE++, have demonstrated comparable and, in indoor scenes, even superior performance to 2D-3D methods, all while maintaining simpler pipeline designs. This highlights the potential of SCR methods to balance high accuracy with computational efficiency, making them increasingly competitive. Nevertheless, 2D-3D localization methods achieve the highest accuracy in structured outdoor environments where dense 3D maps provide a strong reference for precise localization. Despite the success, these methods remain resource-intensive due to their reliance on extensive 3D mapping and feature matching processes.

A key observation from this review is the growing prominence of deep learning-based localization methods across all categories. However, the success of these methods is highly dependent on the availability of large-scale, annotated training data. To address

this challenge, synthetic data generation has emerged as a promising solution. Generative models, such as GAN-based systems, can introduce diverse scenarios like extreme lighting and complex geometry that are difficult to capture in real-world data. However, these synthetic datasets face limitations due to the “domain gap” with real-world data, requiring further research on effective domain adaptation techniques.

The review also highlights several challenges and areas for future research. One critical challenge is improving the generalization capabilities of deep learning-based methods across diverse environments—specifically those that are unseen in training data. To address these issues, future research should prioritize the development of generalizable models that minimize the need for extensive retraining. Scene-agnostic approaches, such as SACReg and PixLoc, demonstrate promising results by leveraging transformer-based architectures and geometric reasoning, but more efficient implementations are needed for real-time deployment. Additionally, multi-sensor fusion offers a robust path forward by integrating visual data with complementary inputs from GPS, IMU, LiDAR, or RADAR. This can provide stability in occluded, texture-less, or extreme-lighting conditions. Future approaches could benefit from improved feature-level and decision-level fusion strategies, particularly using transformer-based architectures for efficient multi-modal data integration, alongside emerging Mamba-style recurrent models for handling spatio-temporal misalignments in dynamic scenes. Lastly, lightweight architectures remain a crucial area of research. Current state-of-the-art models, while accurate, often require substantial computational resources. Developing streamlined networks that maintain high accuracy while reducing resource demands will be essential for real-time applications on resource-constrained devices such as drones and mobile systems. By addressing these research gaps, future studies can unlock new possibilities for more precise, adaptive, and reliable localization systems, with broad implications for both academic research and real-world applications.

Author Contributions: Conceptualization, H.Y., A.L. and H.F.; investigation, H.Y.; writing—original draft preparation, H.Y.; writing and editing, H.Y.; review, H.Y., A.L. and H.F.; supervision, A.L. and H.F.; funding acquisition, A.L.; All authors have read and agreed to the published version of the manuscript.

Funding: Funded by the European Union. Grant number: 101101962-FP6-FutuRe-HORIZON-ER-JU-2022-01. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or Europe’s Rail Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them. The FP6-FutuRe project is supported by Europe’s Rail Joint Undertaking and its members.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Dikmen, M.; Burns, C.M. Autonomous driving in the real world: Experiences with Tesla Autopilot and summon. In Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor, MI, USA, 24–26 October 2016; pp. 225–228.
2. Liu, F.; Lu, Z.; Lin, X. Vision-based environmental perception for autonomous driving. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2025**, *239*, 39–69.
3. Dong, X.; Cappuccio, M.L. Applications of computer vision in autonomous vehicles: Methods, challenges and future directions. *arXiv* **2023**, arXiv:2311.09093.

4. Zhang, K.Z. Applications and prospects of AI in autonomous cars—take Tesla as an example. In Proceedings of the 2nd International Conference on Mechatronic Automation and Electrical Engineering (ICMAEE 2024), Nanjing, China, 22–24 November 2024; pp. 355–360.
5. Wu, Y.; Tang, F.; Li, H. Image-based camera localization: An overview. *Vis. Comput. Ind. Biomed. Art* **2018**, *1*, 8.
6. Piasco, N.; Sidibé, D.; Demonceaux, C.; Gouet-Brunet, V. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognit.* **2018**, *74*, 90–109.
7. Xin, X.; Jiang, J.; Zou, Y. A review of visual-based localization. In Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence, Shanghai, China, 20–22 September 2019; pp. 94–105.
8. Humenberger, M.; Cabon, Y.; Pion, N.; Weinzaepfel, P.; Lee, D.; Guérin, N.; Sattler, T.; Csürka, G. Investigating the role of image retrieval for visual localization: An exhaustive benchmark. *Int. J. Comput. Vis.* **2022**, *130*, 1811–1836.
9. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1470–1477.
10. Nister, D.; Stewenius, H. Scalable recognition with a vocabulary tree. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; pp. 2161–2168.
11. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007.
12. Chum, O.; Philbin, J.; Sivic, J.; Isard, M.; Zisserman, A. Total recall: Automatic query expansion with a generative feature model for object retrieval. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.
13. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665.
14. Perd'och, M.; Chum, O.; Matas, J. Efficient representation of local geometry for large scale object retrieval. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 9–16.
15. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
16. Chum, O.; Mikulík, A.; Perdoch, M.; Matas, J. Total recall II: Query expansion revisited. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 889–896.
17. Arandjelović, R.; Zisserman, A. Three things everyone should know to improve object retrieval. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2911–2918.
18. Torii, A.; Sivic, J.; Pajdla, T.; Okutomi, M. Visual place recognition with repetitive structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2346–2359. <https://doi.org/10.1109/TPAMI.2015.2409868>.
19. Kim, H.J.; Dunn, E.; Frahm, J.M. Predicting good features for image geo-localization using per-bundle vlad. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1170–1178.
20. Torii, A.; Arandjelovic, R.; Sivic, J.; Okutomi, M.; Pajdla, T. 24/7 place recognition by view synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1808–1817.
21. Tolias, G.; Sicre, R.; Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. *arXiv* **2015**, arXiv:1511.05879.
22. Babenko, A.; Lempitsky, V. Aggregating deep convolutional features for image retrieval. *arXiv* **2015**, arXiv:1510.07493.
23. Gordo, A.; Almazán, J.; Revaud, J.; Larlus, D. Deep image retrieval: Learning global representations for image search. In Proceedings of the European Conference on Computer Vision 2016 (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 241–257.
24. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 21–30 June 2016; pp. 5297–5307.
25. Radenović, F.; Tolias, G.; Chum, O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 3–20.
26. Kalantidis, Y.; Mellina, C.; Osindero, S. Cross-dimensional weighting for aggregated deep convolutional features. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 685–701.

27. Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; Han, B. Large-scale image retrieval with attentive deep local features. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3456–3465.
28. Gordo, A.; Almazan, J.; Revaud, J.; Larlus, D. End-to-end learning of deep visual representations for image retrieval. *Int. J. Comput. Vis.* **2017**, *124*, 237–254.
29. Radenović, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1655–1668.
30. Teichmann, M.; Araujo, A.; Zhu, M.; Sim, J. Detect-to-retrieve: Efficient regional aggregation for image search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5109–5118.
31. An, G.; Huo, Y.; Yoon, S.E. Hypergraph propagation and community selection for objects retrieval. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3596–3608.
32. Tolias, G.; Jeníček, T.; Chum, O. Learning and aggregating deep local descriptors for instance-level recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 460–477.
33. Cao, B.; Araujo, A.; Sim, J. Unifying deep local and global features for image search. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 726–743.
34. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14141–14152.
35. Yang, M.; He, D.; Fan, M.; Shi, B.; Xue, X.; Li, F.; Ding, E.; Huang, J. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11772–11781.
36. Wu, H.; Wang, M.; Zhou, W.; Hu, Y.; Li, H. Learning token-based representation for image retrieval. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2703–2711. <https://doi.org/10.1609/aaai.v36i3.20173>.
37. Weinzaepfel, P.; Lucas, T.; Larlus, D.; Kalantidis, Y. Learning super-features for image retrieval. *arXiv* **2022**, arXiv:2201.13182.
38. Shao, S.; Chen, K.; Karpur, A.; Cui, Q.; Araujo, A.; Cao, B. Global features are all you need for image retrieval and reranking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 11036–11046.
39. Tan, F.; Yuan, J.; Ordonez, V. Instance-level image retrieval using reranking transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12105–12115.
40. Wei, T.; Lindenberger, P.; Matas, J.; Barath, D. Breaking the Frame: Visual Place Recognition by Overlap Prediction. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 2322–2331.
41. Mohwald, A.; Jeníček, T.; Chum, O. Dark side augmentation: Generating diverse night examples for metric learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 11153–11163.
42. Zhang, W.; Kosecka, J. Image based localization in urban environments. In Proceedings of the 3rd International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), Chapel Hill, NC, USA, 14–16 June 2006; pp. 33–40.
43. Melekhov, I.; Ylioinas, J.; Kannala, J.; Rahtu, E. Relative camera pose estimation using convolutional neural networks. In Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, Antwerp, Belgium, 18–21 September 2017; pp. 675–687.
44. Laskar, Z.; Melekhov, I.; Kalia, S.; Kannala, J. Camera relocation by computing pairwise relative poses using convolutional neural network. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 929–938.
45. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
46. Balntas, V.; Li, S.; Prisacariu, V. Relocnet: Continuous metric learning relocation using neural nets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 751–767.
47. Saha, S.; Varma, G.; Jawahar, C.V. Improved visual relocation by discovering anchor points. *arXiv* **2018**, arXiv:1811.04370.
48. Ding, M.; Wang, Z.; Sun, J.; Shi, J.; Luo, P. CamNet: Coarse-to-fine retrieval for camera re-localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2871–2880.

49. Li, H.; Zhao, J.; Bazin, J.C.; Chen, W.; Chen, K.; Liu, Y.H. Line-based absolute and relative camera pose estimation in structured environments. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 6914–6920.
50. Zhou, Q.; Sattler, T.; Pollefeys, M.; Leal-Taixe, L. To learn or not to learn: Visual localization from essential matrices. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 3319–3326.
51. Chen, K.; Snavely, N.; Makadia, A. Wide-baseline relative camera pose estimation with directional learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3258–3268.
52. Ullman, S. The interpretation of structure from motion. *Proc. R. Soc. London. Ser. B Biol. Sci.* **1979**, *203*, 405–426.
53. Pan, Y.; Xia, Y.; Li, Y.; Yang, M.; Zhu, Q. Research on stability analysis of large karst cave structure based on multi-source point clouds modeling. *Earth Sci. Inform.* **2023**, *16*, 1637–1656.
54. Tong, X.; Zhang, X.; Liu, S.; Ye, Z.; Feng, Y.; Xie, H.; Chen, L.; Zhang, F.; Han, J.; Jin, Y.; et al. Automatic registration of very low overlapping array InSAR point clouds in urban scenes. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224125.
55. Kabuli, L.A.; Foster, G. Elevation mapping with interferometric synthetic aperture radar for autonomous driving. In Proceedings of the 2024 IEEE Conference on Computational Imaging Using Synthetic Apertures (CISA), Boulder, CO, USA, 20–23 May 2024.
56. da Silva Ruiz, P.R.; Almeida, C.M.; Schimalski, M.B.; Liesenberg, V.; Mitishita, E.A. Multi-approach integration of ALS and TLS point clouds for a 3-D building modeling at LoD3. *Int. J. Archit. Comput.* **2023**, *21*, 652–678.
57. Yang, Y.; Zhao, Z.; Zhou, D.; Lai, Z.; Chang, K.; Fu, T.; Niu, L. Identification and Analysis of the Geohazards Located in an Alpine Valley Based on Multi-Source Remote Sensing Data. *Sensors* **2024**, *24*, 4057.
58. Zhang, W.; Li, Y.; Li, P.; Feng, Z. A BIM and AR-based indoor navigation system for pedestrians on smartphones. *KSCE J. Civ. Eng.* **2025**, *29*, 100005.
59. Wong, M.O.; Lee, S. Indoor navigation and information sharing for collaborative fire emergency response with BIM and multi-user networking. *Autom. Constr.* **2023**, *148*, 104781.
60. Wehbi, R. Integration of BIM and Digital Technologies for Smart Indoor Hazards Management. Ph.D. Thesis, Université de Lille, Lille, France, 2021.
61. Haralick, B.M.; Lee, C.N.; Ottenberg, K.; Nölle, M. Review and analysis of solutions of the three point perspective pose estimation problem. *Int. J. Comput. Vis.* **1994**, *13*, 331–356.
62. Bujnak, M.; Kukulova, Z.; Pajdla, T. A general solution to the P4P problem for camera with unknown focal length. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
63. Bujnak, M.; Kukulova, Z.; Pajdla, T. New efficient solution to the absolute pose problem for camera with unknown focal length and radial distortion. In Proceedings of the Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010; pp. 11–24.
64. Kukulova, Z.; Bujnak, M.; Pajdla, T. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2816–2823.
65. Albl, C.; Kukulova, Z.; Pajdla, T. Rolling shutter absolute pose problem with known vertical direction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2016; pp. 3355–3363.
66. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395.
67. Chum, O.; Matas, J. Optimal randomized RANSAC. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1472–1482.
68. Lebeda, K.; Matas, J.; Chum, O. Fixing the locally optimized ransac—full experimental evaluation. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012.
69. Sattler, T.; Sweeney, C.; Pollefeys, M. On sampling focal length values to solve the absolute pose problem. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 828–843.
70. Barath, D.; Matas, J.; Neskova, J. MAGSAC: Marginalizing sample consensus. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10197–10205.
71. Snavely, N.; Seitz, S.M.; Szeliski, R. Modeling the world from internet photo collections. *Int. J. Comput. Vis.* **2008**, *80*, 189–210.
72. Arth, C.; Wagner, D.; Klopschitz, M.; Irschara, A.; Schmalstieg, D. Wide area localization on mobile phones. In Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality, Orlando, FL, USA, 19–22 October 2009; pp. 73–82.

73. Li, Y.; Snavely, N.; Huttenlocher, D.P. Location recognition using prioritized feature matching. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 791–804.
74. Sattler, T.; Leibe, B.; Kobbelt, L. Fast image-based localization using direct 2D-to-3D matching. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 667–674.
75. Sattler, T.; Leibe, B.; Kobbelt, L. Improving image-based localization by active correspondence search. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 752–765.
76. Paudel, D.P.; Demonceaux, C.; Habed, A.; Vasseur, P. Localization of 2D cameras in a known environment using direct 2D-3D registration. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 196–201.
77. Sattler, T.; Havlena, M.; Radenovic, F.; Schindler, K.; Pollefeys, M. Hyperpoints and fine vocabularies for large-scale location recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2102–2110.
78. Feng, Y.; Fan, L.; Wu, Y. Fast localization in large-scale environments using supervised indexing of binary features. *IEEE Trans. Image Process.* **2015**, *25*, 343–358.
79. Sattler, T.; Leibe, B.; Kobbelt, L. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1744–1756.
80. Liu, L.; Li, H.; Dai, Y. Efficient global 2D-3D matching for camera localization in a large-scale 3D map. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2372–2381.
81. Song, Z.; Wang, C.; Liu, Y.; Shen, S. Recalling direct 2D-3D matches for large-scale visual localization. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 1191–1197.
82. Nadeem, U.; Jalwana, M.A.; Bennamoun, M.; Togneri, R.; Sohel, F. Direct image to point cloud descriptors matching for 6-dof camera localization in dense 3D point clouds. In Proceedings of the International Conference on Neural Information Processing, Vancouver, BC, Canada, 8–14 December 2019; pp. 222–234.
83. Nadeem, U.; Bennamoun, M.; Togneri, R.; Sohel, F. Unconstrained Matching of 2D and 3D Descriptors for 6-DOF Pose Estimation. *arXiv* **2020**, arXiv:2005.14502.
84. Feng, M.; Hu, S.; Ang, M.H.; Lee, G.H. 2D3D-MatchNet: Learning to match keypoints across 2D image and 3D point cloud. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 4790–4796.
85. Pham, Q.H.; Uy, M.A.; Hua, B.S.; Nguyen, D.T.; Roig, G.; Yeung, S.K. LCD: Learned cross-domain descriptors for 2D-3D matching. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11856–11864.
86. Yu, H.; Ye, W.; Feng, Y.; Bao, H.; Zhang, G. Learning bipartite graph matching for robust visual localization. In Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Recife/Porto de Galinhas, Brazil, 9–13 November 2020; pp. 146–155.
87. Yu, H.; Zhen, W.; Yang, W.; Zhang, J.; Scherer, S. Monocular camera localization in prior lidar maps with 2D-3D line correspondences. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 4588–4594.
88. Sarlin, P.E.; Unagar, A.; Larsson, M.; Germain, H.; Toft, C.; Larsson, V.; Pollefeys, M.; Lepetit, V.; Hammarstrand, L.; Kahl, F.; et al. Back to the feature: Learning robust camera localization from pixels to pose. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3247–3257.
89. Lai, B.; Liu, W.; Wang, C.; Fan, X.; Lin, Y.; Bian, X.; Wu, S.; Cheng, M.; Li, J. 2D3D-MVPNet: Learning cross-domain feature descriptors for 2D-3D matching based on multi-view projections of point clouds. *Appl. Intell.* **2022**, *52*, 14178–14193.
90. Kim, M.; Koo, J.; Kim, G. Ep2p-loc: End-to-end 3D point to 2D pixel localization for large-scale visual localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 21527–21537.
91. Zhou, Q.; Agostinho, S.; Ošep, A.; Leal-Taixé, L. Is geometry enough for matching in visual localization? In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 407–425.
92. Nguyen, S.T.; Fontan, A.; Milford, M.; Fischer, T. FUSELOC: Fusing Global and Local Descriptors to Disambiguate 2D-3D Matching in Visual Localization. *arXiv* **2024**, arXiv:2408.12037.

93. Irschara, A.; Zach, C.; Frahm, J.M.; Bischof, H. From structure-from-motion point clouds to fast location recognition. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2599–2606.
94. Sattler, T.; Weyand, T.; Leibe, B.; Kobbelt, L. Image retrieval for image-based localization revisited. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012; p. 4.
95. Cao, S.; Snavely, N. Minimal scene descriptions from structure from motion models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 461–468.
96. Sattler, T.; Torii, A.; Sivic, J.; Pollefeys, M.; Taira, H.; Okutomi, M.; Pajdla, T. Are large-scale 3D models really necessary for accurate visual localization? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1637–1646.
97. Camposeco, F.; Cohen, A.; Pollefeys, M.; Sattler, T. Hybrid camera pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 136–144.
98. Taira, H.; Okutomi, M.; Sattler, T.; Cimpoi, M.; Pollefeys, M.; Sivic, J.; Pajdla, T.; Torii, A. InLoc: Indoor visual localization with dense matching and view synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7199–7209.
99. Sarlin, P.E.; Debraine, F.; Dymczyk, M.; Siegwart, R.; Cadena, C. Leveraging deep visual descriptors for hierarchical efficient localization. In Proceedings of the Conference on Robot Learning, Zurich, Switzerland, 29–31 October 2018; pp. 456–465.
100. Sarlin, P.E.; Cadena, C.; Siegwart, R.; Dymczyk, M. From coarse to fine: Robust hierarchical localization at large scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12716–12725.
101. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-net: A trainable CNN for joint description and detection of local features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8092–8101.
102. Germain, H.; Bourmaud, G.; Lepetit, V. S2dnet: Learning accurate correspondences for sparse-to-dense feature matching. *arXiv* **2020**, arXiv:2004.01673.
103. Yang, T.Y.; Nguyen, D.K.; Heijnen, H.; Balntas, V. Ur2kid: Unifying retrieval, keypoint detection, and keypoint description without local correspondence supervision. *arXiv* **2020**, arXiv:2001.07252.
104. Shi, T.; Cui, H.; Song, Z.; Shen, S. Dense semantic 3D map based long-term visual localization with hybrid features. *arXiv* **2020**, arXiv:2005.10766.
105. Humenberger M, Cabon Y, Guerin N, Morat J, Leroy V, Revaud J, Rerole P, Pion N, De Souza C, Csurka G. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*. 2020 Jul 27.
106. Shu, M.; Chen, G.; Zhang, Z. Efficient image-based indoor localization with MEMS aid on the mobile device. *ISPRS J. Photogramm. Remote Sens.* **2022**, *185*, 85–110.
107. Yan, S.; Liu, Y.; Wang, L.; Shen, Z.; Peng, Z.; Liu, H.; Zhang, M.; Zhang, G.; Zhou, X. Long-term visual localization with mobile sensors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 July 2023; pp. 17245–17255.
108. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
109. Kendall, A.; Cipolla, R. Modelling uncertainty in deep learning for camera relocalization. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 4762–4769.
110. Kendall, A.; Cipolla, R. Geometric loss functions for camera pose regression with deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5974–5983.
111. Walch, F.; Hazirbas, C.; Leal-Taixe, L.; Sattler, T.; Hilsenbeck, S.; Cremers, D. Image-based localization using LSTMs for structured feature correlation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 627–637.
112. Melekhov, I.; Ylioinas, J.; Kannala, J.; Rahtu, E. Image-based localization using hourglass networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 879–886.
113. Wu, J.; Ma, L.; Hu, X. Delving deeper into convolutional neural networks for camera relocalization. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 5644–5651.

114. Naseer, T.; Burgard, W. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 1525–1530.
115. Brahmabhatt, S.; Gu, J.; Kim, K.; Hays, J.; Kautz, J. Geometry-aware learning of maps for camera localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2616–2625.
116. Wang, B.; Chen, C.; Lu, C.X.; Zhao, P.; Trigoni, N.; Markham, A. AtLoc: Attention guided camera localization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 10393–10401.
117. Cai, M.; Shen, C.; Reid, I. A Hybrid Probabilistic Model for Camera Relocalization. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; p. 8.
118. Chidlovskii, B.; Sadek, A. Adversarial transfer of pose estimation regression. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 646–661.
119. Shavit, Y.; Ferens, R. Do we really need scene-specific pose encoders? In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3186–3192.
120. Blanton, H.; Greenwell, C.; Workman, S.; Jacobs, N. Extending absolute pose regression to multiple scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 38–39.
121. Shavit, Y.; Ferens, R.; Keller, Y. Learning multi-scene absolute pose regression with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2733–2742.
122. Shavit, Y.; Keller, Y. Camera pose auto-encoders for improving pose regression. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 140–157.
123. Clark, R.; Wang, S.; Markham, A.; Trigoni, N.; Wen, H. VidLoc: A deep spatio-temporal model for 6-dof video-clip relocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6856–6864.
124. Valada, A.; Radwan, N.; Burgard, W. Deep auxiliary learning for visual localization and odometry. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 6939–6946.
125. Radwan, N.; Valada, A.; Burgard, W. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robot. Autom. Lett.* **2018**, *3*, 4407–4414.
126. Bui, M.; Baur, C.; Navab, N.; Ilic, S.; Albarqouni, S. Adversarial networks for camera pose regression and refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
127. Wang, S.; Kang, Q.; She, R.; Tay, W.P.; Hartmannsgruber, A.; Navarro, D.N. RobustLoc: Robust camera pose regression in challenging driving environments. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2023; pp. 6209–6216.
128. Xu, M.; Zhang, Z.; Gong, Y.; Poslad, S. Regression-based camera pose estimation through multi-level local features and global features. *Sensors* **2023**, *23*, 4063.
129. Chen, S.; Li, X.; Wang, Z.; Prisacariu, V.A. Dfnet: Enhance absolute pose regression with direct feature matching. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 1–17.
130. Chen, S.; Bhalgat, Y.; Li, X.; Bian, J.W.; Li, K.; Wang, Z.; Prisacariu, V.A. Neural refinement for absolute pose regression with feature synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 20987–20996.
131. Shotton, J.; Glocker, B.; Zach, C.; Izadi, S.; Criminisi, A.; Fitzgibbon, A. Scene coordinate regression forests for camera relocalization in RGB-D images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2930–2937.
132. Guzman-Rivera, A.; Kohli, P.; Glocker, B.; Shotton, J.; Sharp, T.; Fitzgibbon, A.; Izadi, S. Multi-output learning for camera relocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1114–1121.
133. Valentin, J.; Nießner, M.; Shotton, J.; Fitzgibbon, A.; Izadi, S.; Torr, P.H. Exploiting uncertainty in regression forests for accurate camera relocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4400–4408.

134. Brachmann, E.; Krull, A.; Nowozin, S.; Shotton, J.; Michel, F.; Gumhold, S.; Rother, C. DSAC—Differentiable RANSAC for camera localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6684–6692.
135. Brachmann, E.; Rother, C. Learning less is more-6D camera localization via 3D surface regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4654–4662.
136. Brachmann, E.; Rother, C. Expert sample consensus applied to camera re-localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7525–7534.
137. Brachmann, E.; Rother, C. Visual camera re-localization from RGB and RGB-D images using DSAC. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5847–5865.
138. Li, X.; Wang, S.; Zhao, Y.; Verbeek, J.; Kannala, J. Hierarchical scene coordinate classification and regression for visual localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11983–11992.
139. Wang, S.; Laskar, Z.; Melekhov, I.; Li, X.; Zhao, Y.; Tolas, G.; Kannala, J. HSCNet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *Int. J. Comput. Vis.* **2024**, *132*, 2530–2550.
140. Rekavandi, A.M.; Boussaid, F.; Seghouane, A.-K.; Bennamoun, M. B-Pose: Bayesian Deep Network for Accurate Camera 6-DoF Pose Estimation from RGB Images. *IEEE Robot. Autom. Lett.* **2023**, *8*, 6746–6754.
141. Tang, S.; Tang, S.; Tagliasacchi, A.; Tan, P.; Furukawa, Y. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 July 2023; pp. 929–939.
142. Chen, S.; Cavallari, T.; Prisacariu, V.A.; Brachmann, E. Map-relative pose regression for visual re-localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 20665–20674.
143. Revaud, J.; Cabon, Y.; Brégier, R.; Lee, J.; Weinzaepfel, P. Sacreg: Scene-agnostic coordinate regression for visual localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 688–698.
144. Brachmann, E.; Cavallari, T.; Prisacariu, V.A. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 July 2023; pp. 5044–5053.
145. Lu, D.; Xiao, W.; Ran, T.; Yuan, L.; Lv, K.; Zhang, J. Attention-Based Accelerated Coordinate Encoding Network for Visual Relocalization. In Proceedings of the 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 20–22 September 2024; pp. 1675–1680.
146. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
147. Jegou, H.; Douze, M.; Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 304–317.
148. Radenović, F.; Iscen, A.; Tolas, G.; Avrithis, Y.; Chum, O. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5706–5715.
149. Weyand, T.; Araujo, A.; Cao, B.; Sim, J. Google Landmarks Dataset V2—A large-scale benchmark for instance-level recognition and retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2575–2584.
150. Badino, H.; Huber, D.; Kanade, T. Visual topometric localization. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 5–9 June 2011; pp. 794–799.
151. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The Oxford RobotCar dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15.
152. Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The ApolloScape dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 954–960.
153. Sattler, T.; Maddern, W.; Toft, C.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; et al. Benchmarking 6DoF outdoor visual localization in changing conditions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8601–8610.

154. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
155. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
156. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, online, 6 November 2020; pp. 1597–1607.
157. Liu, Y.; Dong, Q. EquiPose: Exploiting Permutation Equivariance for Relative Camera Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 1127–1137.
158. Ferens, R.; Keller, Y. HyperPose: Hypernetwork-infused camera pose localization and an extended cambridge landmarks dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025), Nashville, TN, USA, 11–15 June 2025; pp. 11547–11557.
159. Dong, S.; Wang, S.; Liu, S.; Cai, L.; Fan, Q.; Kannala, J.; Yang, Y. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025), Nashville, TN, USA, 11–15 June 2025; pp. 16739–16752.
160. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.; Han, S. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird’s-Eye View Representation. *arXiv* **2022**, arXiv:2205.13542.
161. Palladin, E.; Dietze, R.; Narayanan, P.; Bijelic, M.; Heide, F. SAMFusion: Sensor-adaptive multimodal fusion for 3D object detection in adverse weather. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 484–503.
162. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the 1st Conference on Language Modeling, Philadelphia, PA, USA, 11 April–10 May 2024.
163. Dao, T.; Gu, A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv* **2024**, arXiv:2405.21060.
164. Tang, Y.; Dong, P.; Tang, Z.; Chu, X.; Liang, J. VMRNN: Integrating Vision Mamba and LSTM for efficient and accurate spatio-temporal forecasting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 5663–5673.
165. Yan, W.; Yin, F.; Wang, J.; Leus, G.; Zoubir, A.M.; Tian, Y. Attentional Graph Neural Network Is All You Need for Robust Massive Network Localization. *arXiv* **2023**, arXiv:2311.16856.
166. Huang, J.; Wu, M.; Li, P.; Wu, W.; Yu, R. VimGeo: Efficient Cross-View Geo-Localization with Vision Mamba Architecture. In Proceedings of the 34th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 16–22 August 2025; pp. 1188–1196.
167. Hong, C.Y.; Wang, L.H.; Liu, T.L. Promptable 3-D Object Localization with Latent Diffusion Models. In Proceedings of the 39th Annual Conference on Neural Information Processing Systems, San Diego, CA, 2–7 December 2025.
168. Xu, Q.; Chen, Y.; Li, Y.; Liu, Z.; Lou, Z.; Zhang, Y.; Zheng, H.; He, X. MambaVesselNet++: A hybrid CNN-Mamba architecture for medical image segmentation. *arXiv* **2025**, arXiv:2507.19931.
169. Boukhari, D.E. Mamba-CNN: A Hybrid Architecture for Efficient and Accurate Facial Beauty Prediction. *arXiv* **2025**, arXiv:2509.01431.
170. Cao, A.; Li, Z.; Jomsky, J.; Laine, A.F.; Guo, J. MedSegMamba: 3D CNN-Mamba hybrid architecture for brain segmentation. *arXiv* **2024**, arXiv:2409.08307.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.